Final Project – Milestone 1

Mohammed Saif Wasay (002815958)

Masters of Professional Studies in Informatics, Northeastern University

ALY 6010: Probability Theory and Introductory Statistics

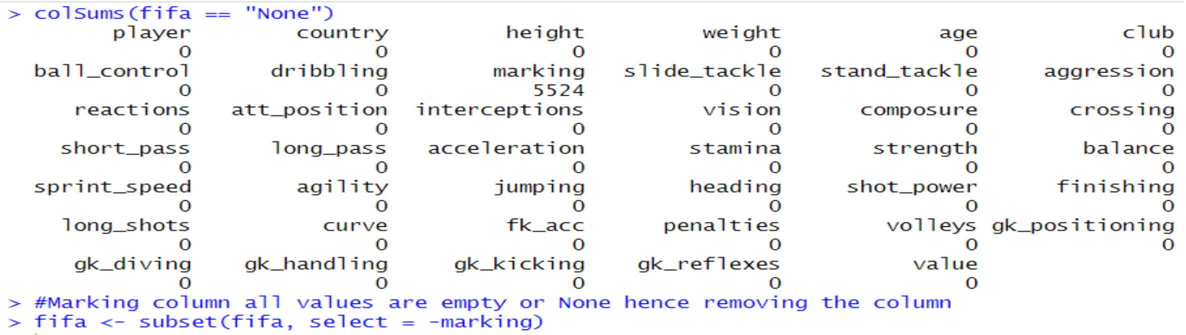
Dr. Harpreet Sharma

01st June 2024

Introduction

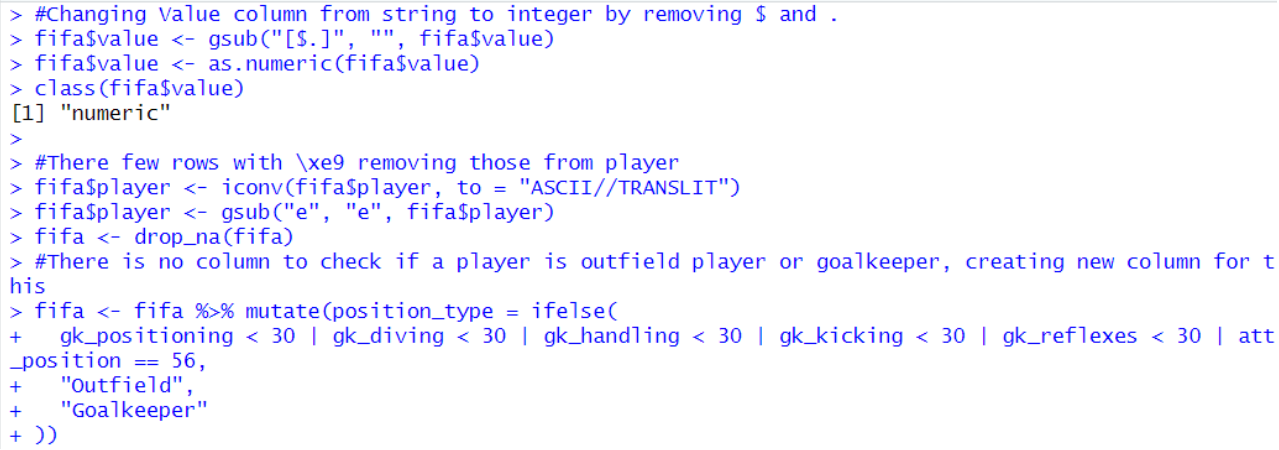
In this project, In this project, we will delve into the FIFA Football Players Dataset, which offers a wealth of information about football players worldwide. This dataset is rich in both numerical and textual data, covering various aspects of each player. It includes statistics for 5,682 players across 41 columns, with details such as country, height, weight, dribbling skills, attacking position, goalkeeping abilities (like positioning and diving), and player value. This dataset is an excellent resource for conducting in-depth analyses and gaining insights into the world of football, appealing to both gaming enthusiasts and real-world sports fans.

The dataset contains four categorical columns: Player, Country, Club, and Value. The remaining columns are numerical. To start the data cleaning process, we first examine and address any null values. We remove the "marking" column because it is almost entirely filled with null or "None" values. Additionally, to make the "Value" column usable for analysis, we convert it to numeric format by extracting only the integer data from the column.



```
> colSums(fifa == "None")
        player        country         height         weight            age           club 
             0              0              0              0              0              0 
  ball_control      dribbling        marking    slide_tackle   stand_tackle     aggression 
             0              0           5524              0              0              0 
     reactions   att_position  interceptions         vision      composure       crossing 
             0              0              0              0              0              0 
    short_pass      long_pass   acceleration        stamina       strength        balance 
             0              0              0              0              0              0 
  sprint_speed        agility        jumping        heading     shot_power      finishing 
             0              0              0              0              0              0 
    long_shots          curve         fk_acc      penalties        volleys gk_positioning 
             0              0              0              0              0              0 
     gk_diving    gk_handling     gk_kicking    gk_reflexes          value 
             0              0              0              0              0 
> #Marking column all values are empty or None hence removing the column
> fifa <- subset(fifa, select = -marking)
```

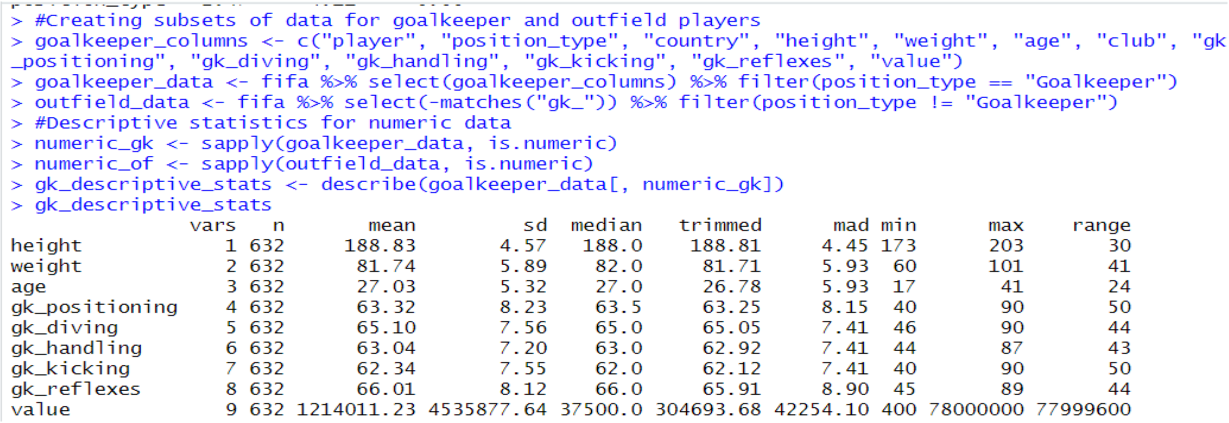
Some players had ASCII characters in their names, which led to duplicate rows in the dataset. To address this, we removed the duplicates. Additionally, we introduced a new column called "position_type" to indicate whether a player is a goalkeeper or an outfield player.



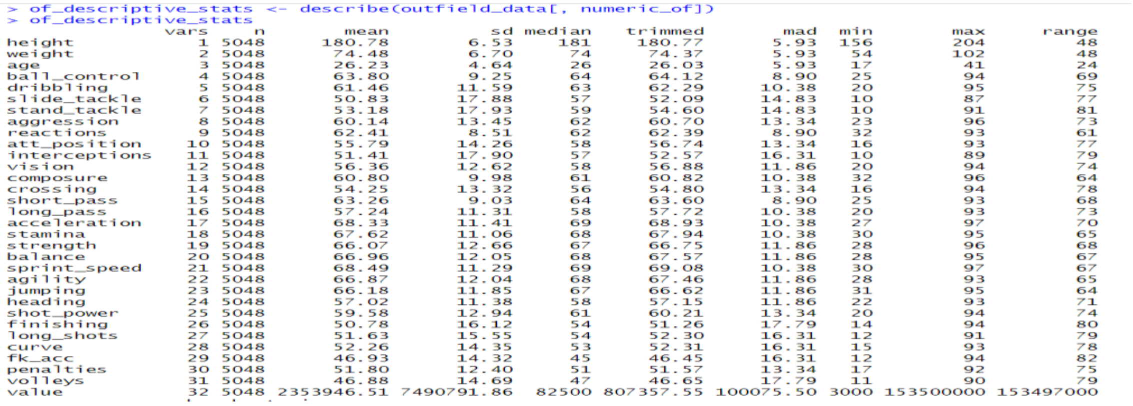
```
> #Changing Value column from string to integer by removing $ and .
> fifa$value <- gsub("[$.]", "", fifa$value)
> fifa$value <- as.numeric(fifa$value)
> class(fifa$value)
[1] "numeric"
> 
> #There few rows with \xe9 removing those from player
> fifa$player <- iconv(fifa$player, to = "ASCII//TRANSLIT")
> fifa$player <- gsub("e", "e", fifa$player)
> fifa <- drop_na(fifa)
> #There is no column to check if a player is outfield player or goalkeeper, creating new column for t
his
> fifa <- fifa %>% mutate(position_type = ifelse(
+   gk_positioning < 30 | gk_diving < 30 | gk_handling < 30 | gk_kicking < 30 | gk_reflexes < 30 | att
_position == 56,
+   "Outfield",
+   "Goalkeeper"
+ ))
```

Data Analysis

Thus, subsets of the data are created to focus specifically on goalkeepers and outfield players. For the goalkeepers, we included only the relevant variables, such as goalkeeping-specific stats. Similarly, for the outfield players, we excluded any attributes related to goalkeeping. After organizing these subsets, we generated descriptive statistics for each dataset, including measures like mean, median, standard deviation, minimum, and maximum values, among others.

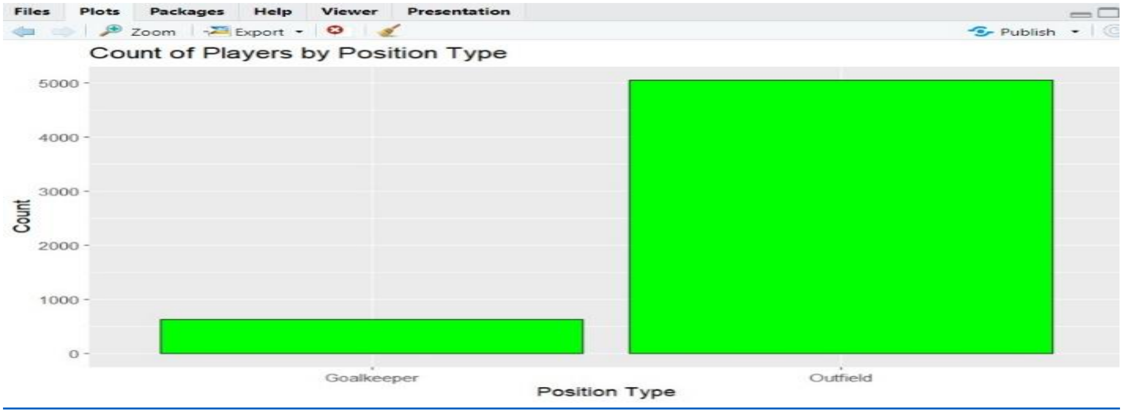


```
> #Creating subsets of data for goalkeeper and outfield players
> goalkeeper_columns <- c("player", "position_type", "country", "height", "weight", "age", "club", "gk
_positioning", "gk_diving", "gk_handling", "gk_kicking", "gk_reflexes", "value")
> goalkeeper_data <- fifa %>% select(goalkeeper_columns) %>% filter(position_type == "Goalkeeper")
> outfield_data <- fifa %>% select(-matches("gk_")) %>% filter(position_type != "Goalkeeper")
> #Descriptive statistics for numeric data
> numeric_gk <- sapply(goalkeeper_data, is.numeric)
> numeric_of <- sapply(outfield_data, is.numeric)
> gk_descriptive_stats <- describe(goalkeeper_data[, numeric_gk])
> gk_descriptive_stats
               vars   n       mean        sd  median   trimmed      mad min      max    range
height            1 632     188.83      4.57   188.0    188.81     4.45 173      203       30
weight            2 632      81.74      5.89    82.0     81.71     5.93  60      101       41
age               3 632      27.03      5.32    27.0     26.78     5.93  17       41       24
gk_positioning    4 632      63.32      8.23    63.5     63.25     8.15  40       90       50
gk_diving         5 632      65.10      7.56    65.0     65.05     7.41  46       90       44
gk_handling       6 632      63.04      7.20    63.0     62.92     7.41  44       87       43
gk_kicking        7 632      62.34      7.55    62.0     62.12     7.41  40       90       50
gk_reflexes       8 632      66.01      8.12    66.0     65.91     8.90  45       89       44
value             9 632 1214011.23 4535877.64 37500.0 304693.68 42254.10 400 78000000 77999600
```

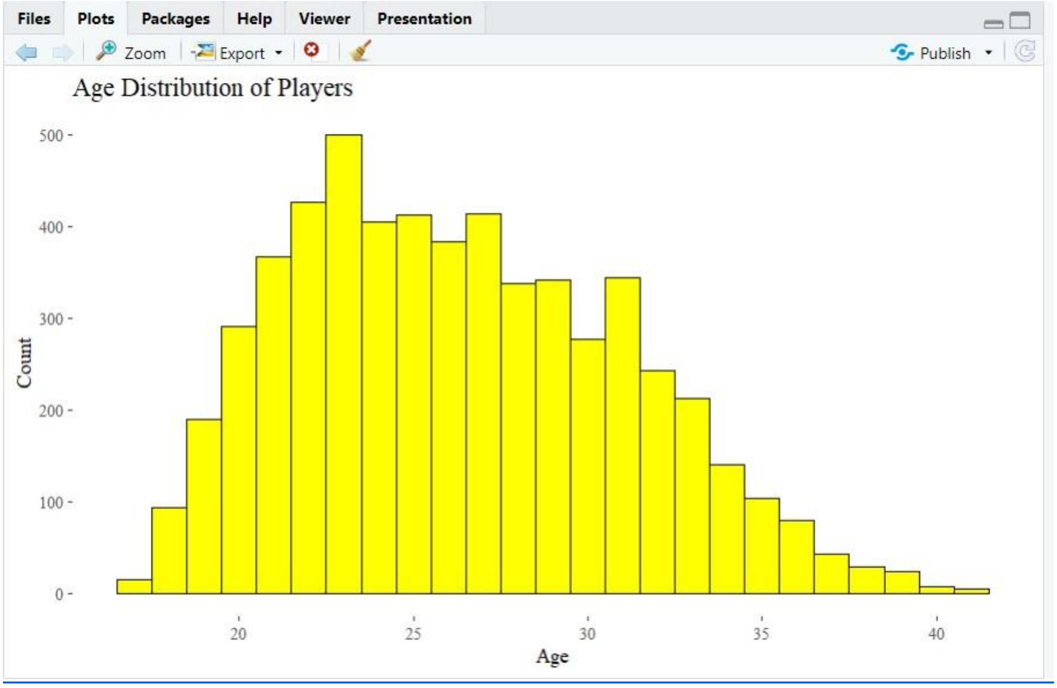


```
> of_descriptive_stats <- describe(outfield_data[, numeric_of])
> of_descriptive_stats
              vars    n       mean         sd median   trimmed       mad  min       max     range
height           1 5048     180.78       6.53    181    180.77      5.93  156       204        48
weight           2 5048      74.48       6.70     74     74.37      5.93   54       102        48
age              3 5048      26.23       4.64     26     26.03      5.93   17        41        24
ball_control     4 5048      63.80       9.25     64     64.12      8.90   25        94        69
dribbling        5 5048      61.46      11.59     63     62.29     10.38   20        95        75
slide_tackle     6 5048      50.83      17.88     57     52.09     14.83   10        87        77
stand_tackle     7 5048      53.18      17.93     59     54.60     14.83   10        91        81
aggression       8 5048      60.14      13.45     62     60.70     13.34   23        96        73
reactions        9 5048      62.41       8.51     62     62.39      8.90   32        93        61
att_position    10 5048      55.79      14.26     58     56.74     13.34   16        93        77
interceptions   11 5048      51.41      17.90     57     52.57     16.31   10        89        79
vision          12 5048      56.36      12.62     58     56.88     11.86   20        94        74
composure       13 5048      60.80       9.98     61     60.82     10.38   32        96        64
crossing        14 5048      54.25      13.32     56     54.80     13.34   16        94        78
short_pass      15 5048      63.26       9.03     64     63.60      8.90   25        93        68
long_pass       16 5048      57.24      11.31     58     57.72     10.38   20        93        73
acceleration    17 5048      68.33      11.41     69     68.93     10.38   27        97        70
stamina         18 5048      67.62      11.06     68     67.94     10.38   30        95        65
strength        19 5048      66.07      12.66     67     66.75     11.86   28        96        68
balance         20 5048      66.96      12.05     68     67.57     11.86   28        95        67
sprint_speed    21 5048      68.49      11.29     69     69.08     10.38   30        97        67
agility         22 5048      66.87      12.04     68     67.46     11.86   28        93        65
jumping         23 5048      66.18      11.85     67     66.62     11.86   31        95        64
heading         24 5048      57.02      11.38     58     57.15     11.86   22        93        71
shot_power      25 5048      59.58      12.94     61     60.21     13.34   20        94        74
finishing       26 5048      50.78      16.12     54     51.26     17.79   14        94        80
long_shots      27 5048      51.63      15.55     54     52.30     16.31   12        91        79
curve           28 5048      52.26      14.35     53     52.31     16.31   15        93        78
fk_acc          29 5048      46.93      14.32     45     46.45     16.31   12        94        82
penalties       30 5048      51.80      12.40     51     51.57     13.34   17        92        75
volleys         31 5048      46.88      14.69     47     46.65     17.79   11        90        79
value           32 5048 2353946.51 7490791.86  82500 807357.55 100075.50 3000 153500000 153497000
```

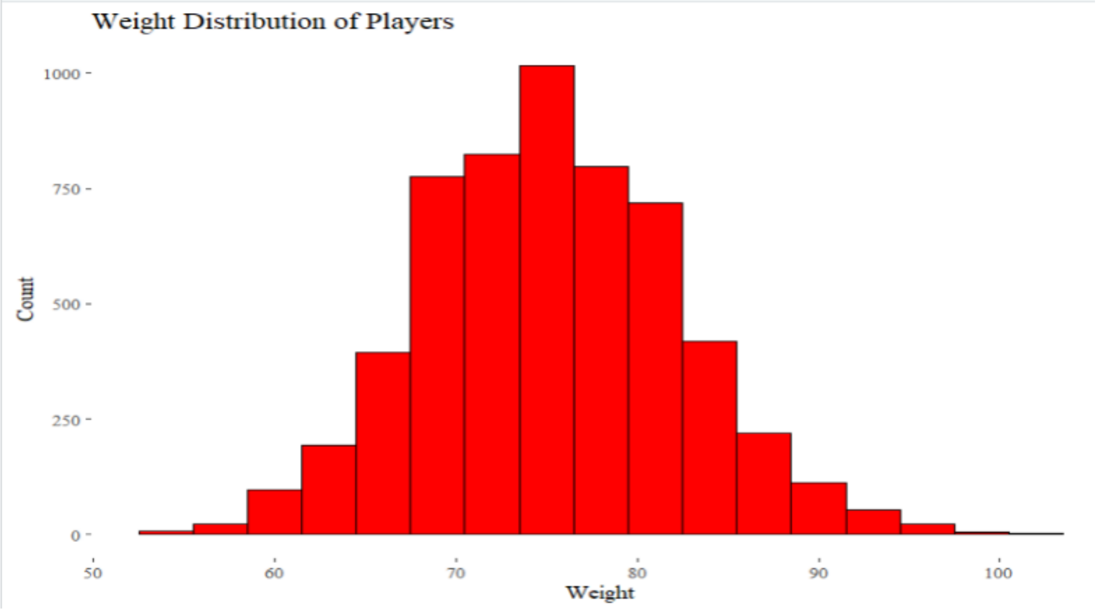
Visualized the distribution of outfield players and goalkeepers in the dataset. The visualizations clearly show that the majority of the players are outfielders.



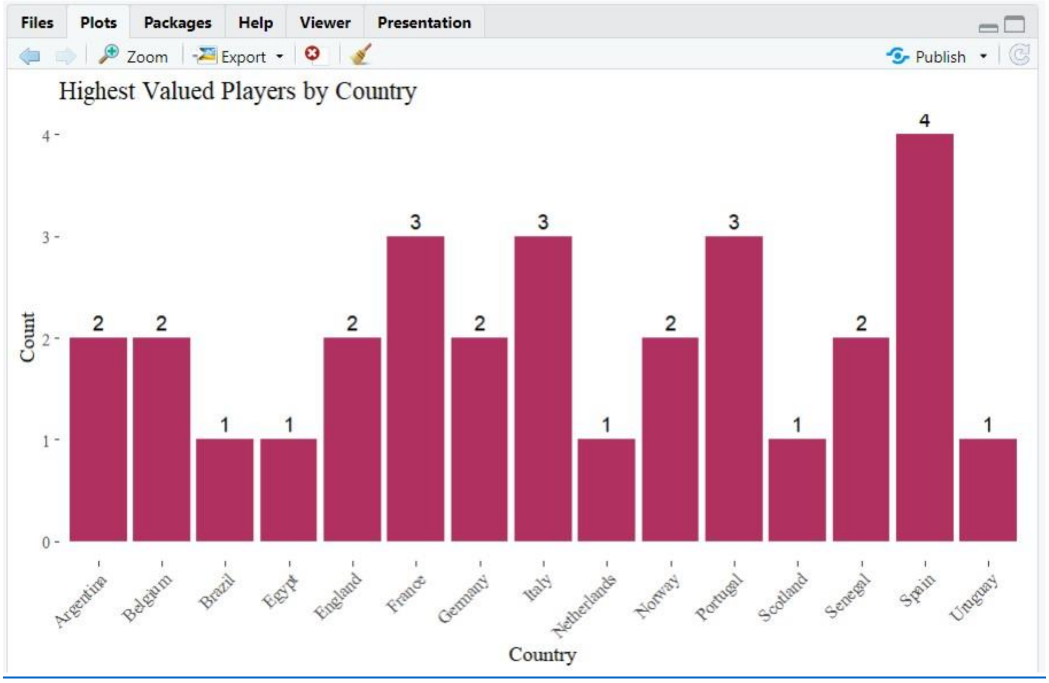
Created a histogram to analyze the age distribution of players in the dataset. This visualization revealed that most players are between the ages of 22 and 27, with the highest concentration around age 23, which includes approximately 500 individuals.



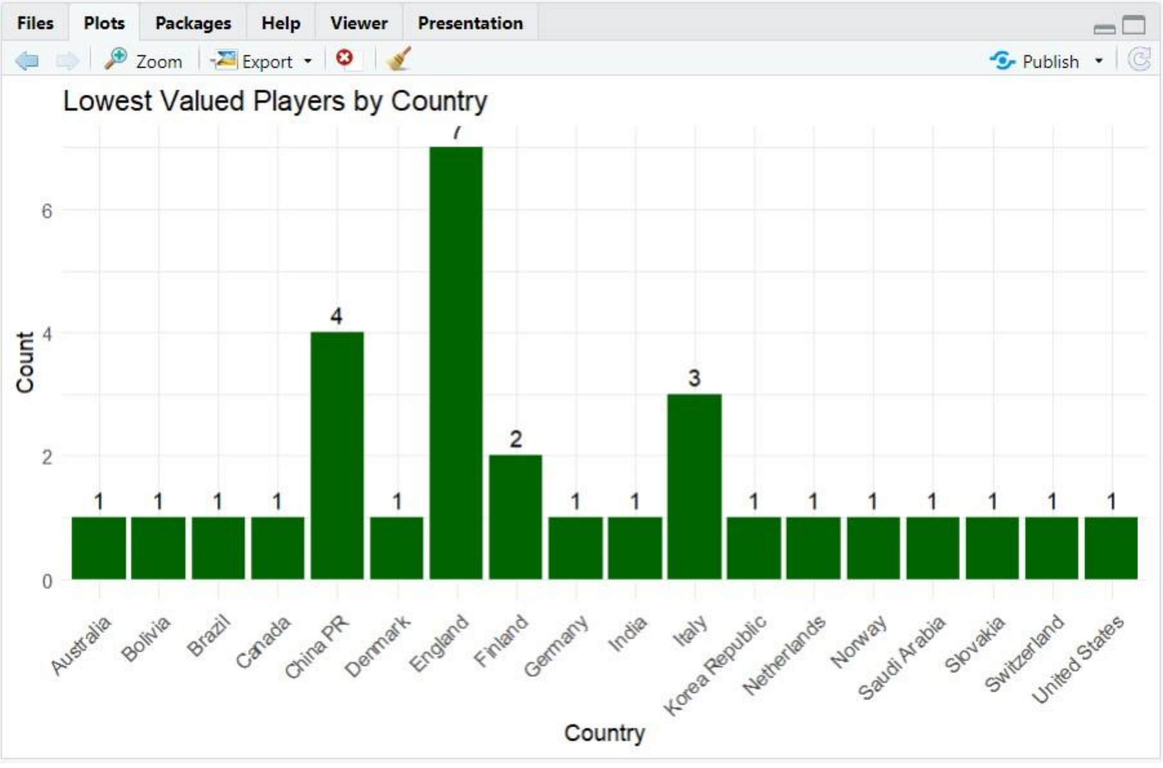
Visualized the weight distribution of the players and discovered that it follows the Normal Distribution.



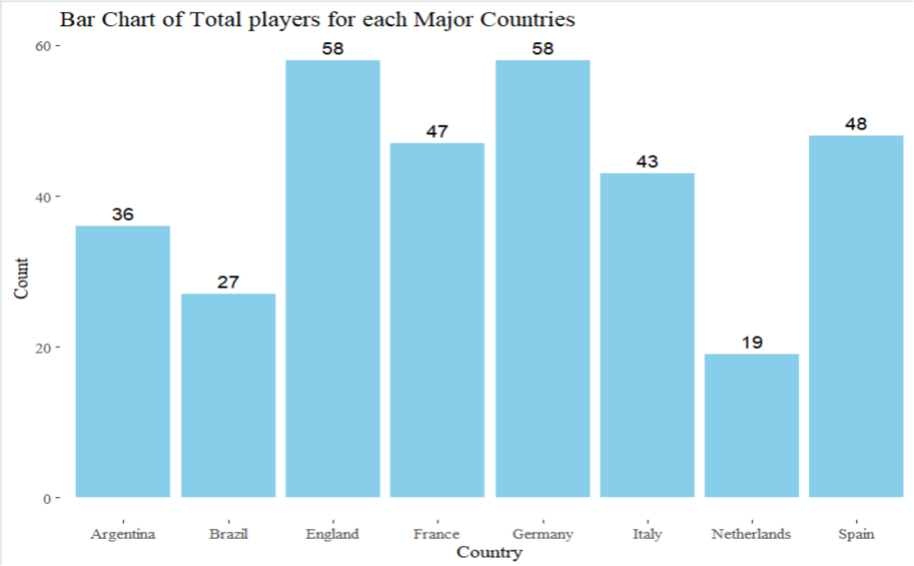
We compiled a list of the 30 most and 30 least valuable players in the dataset. When examining this data by player nationality, we found that Spain has the highest number of players in the top 30 valued players, with four representatives. This is followed by Portugal, Italy, and France, each with three players in the top 30.



Similarly, when it comes to the lowest valued players, we can see that 7 of the 30 lowest valued players are English.

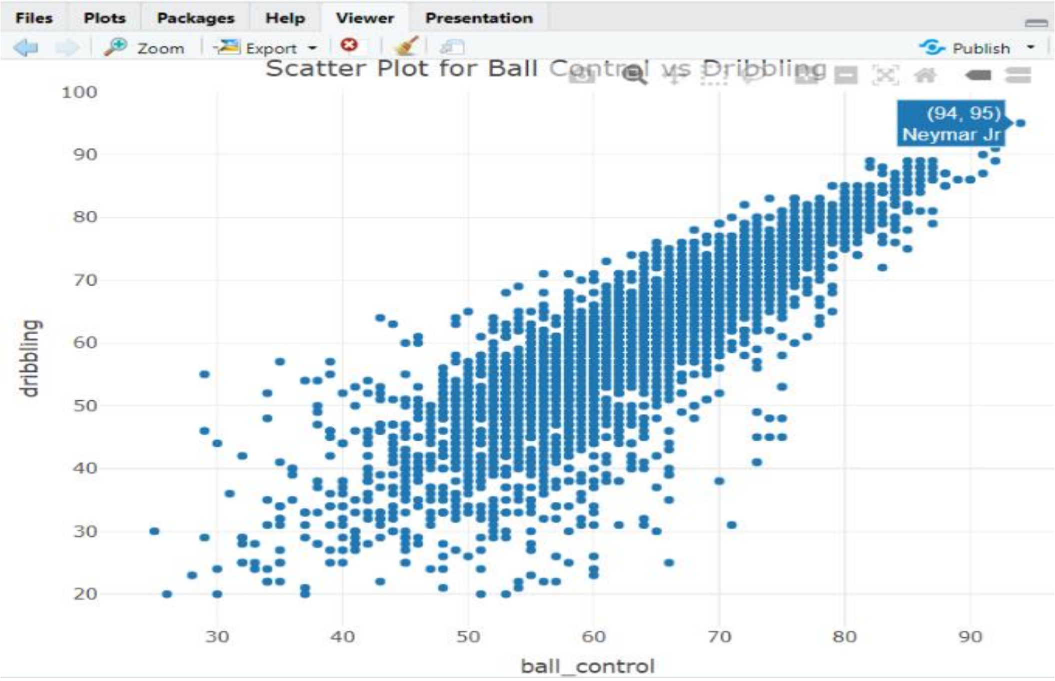


We created a subset of the dataset focusing on players from key footballing nations. Upon visualizing this data, we found that Germany and England have the highest number of players, each with 58, followed by Spain and France.

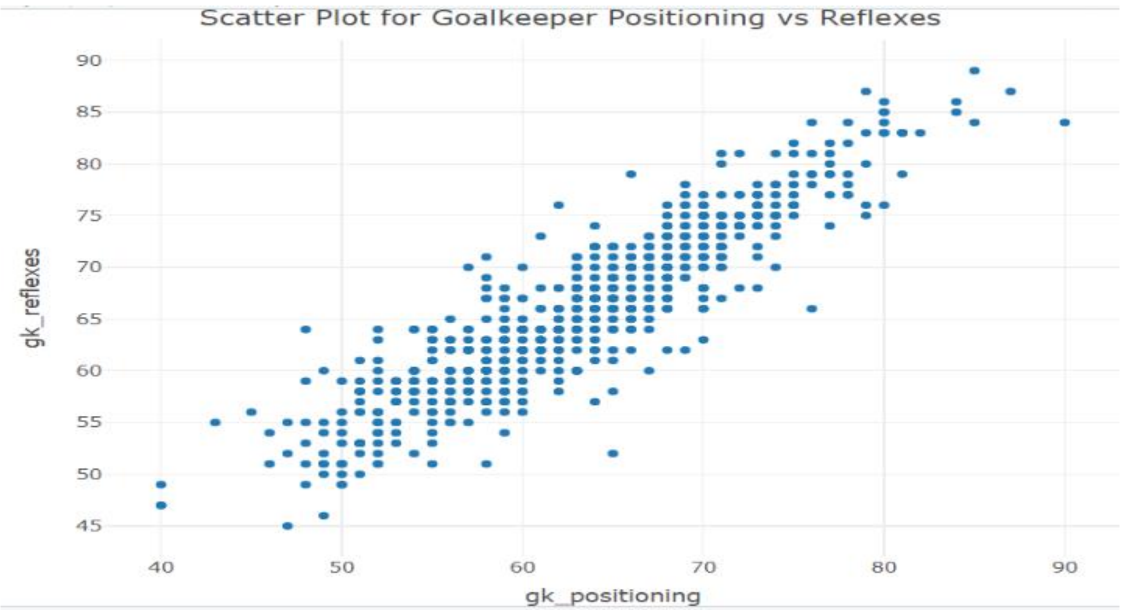


Created an interactive scatter plot to visualize the relationship between players' ball control and dribbling abilities.

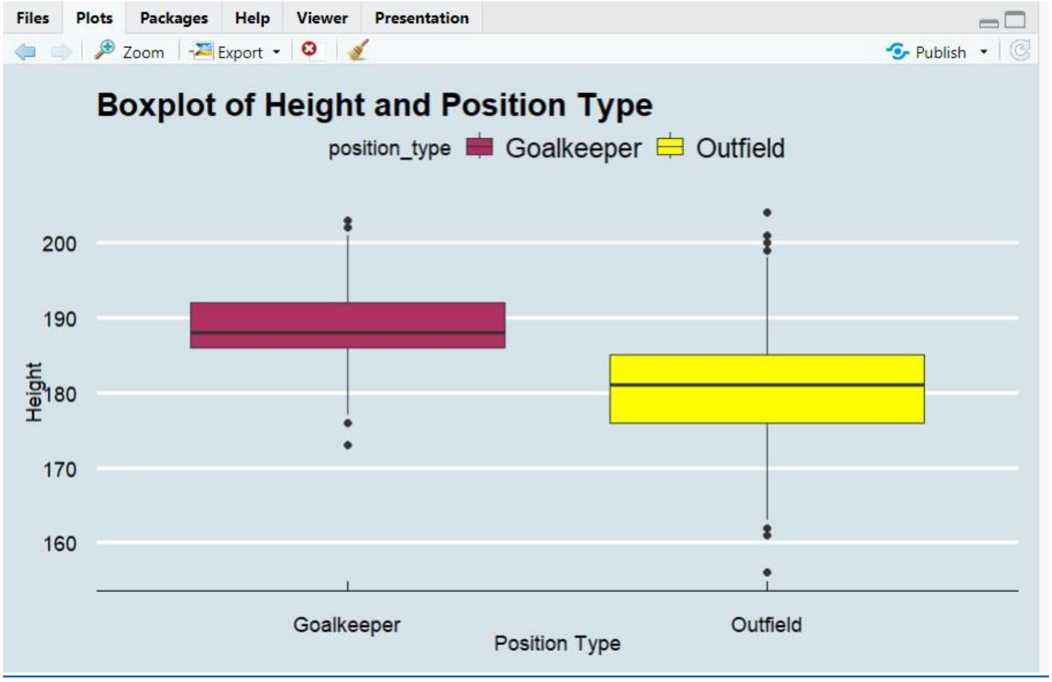
This plot revealed a positive correlation, indicating that players with better ball control tend to have better dribbling skills.

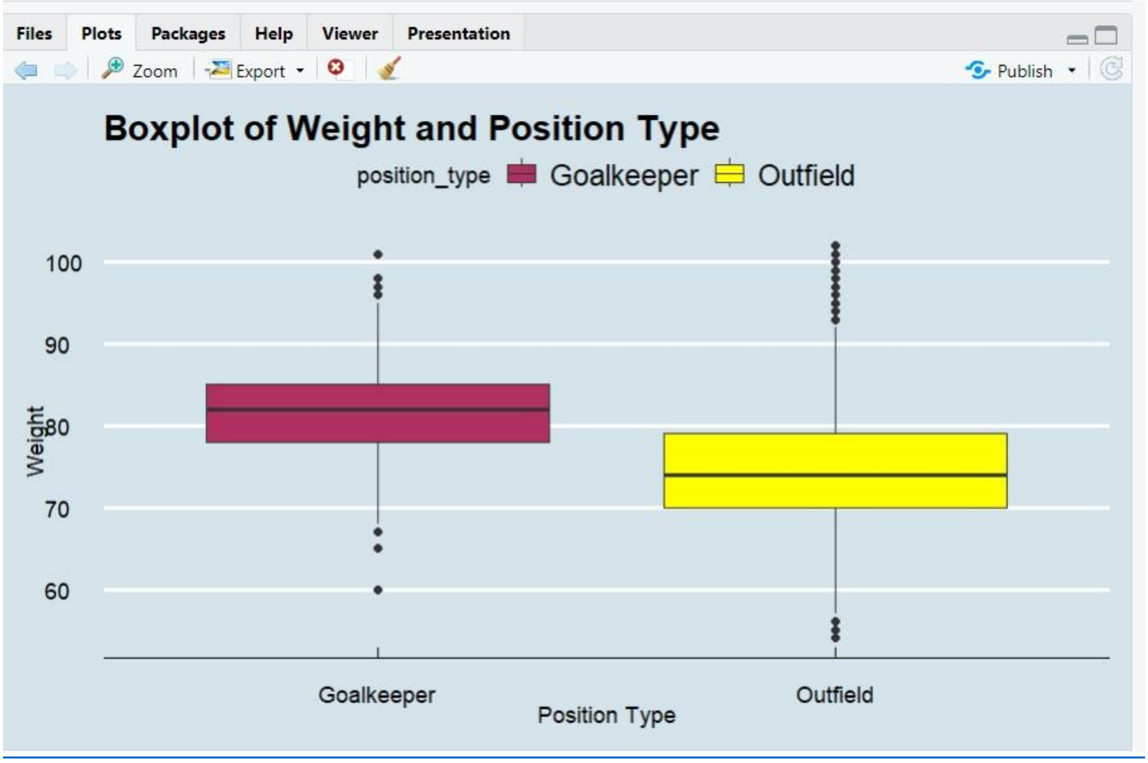


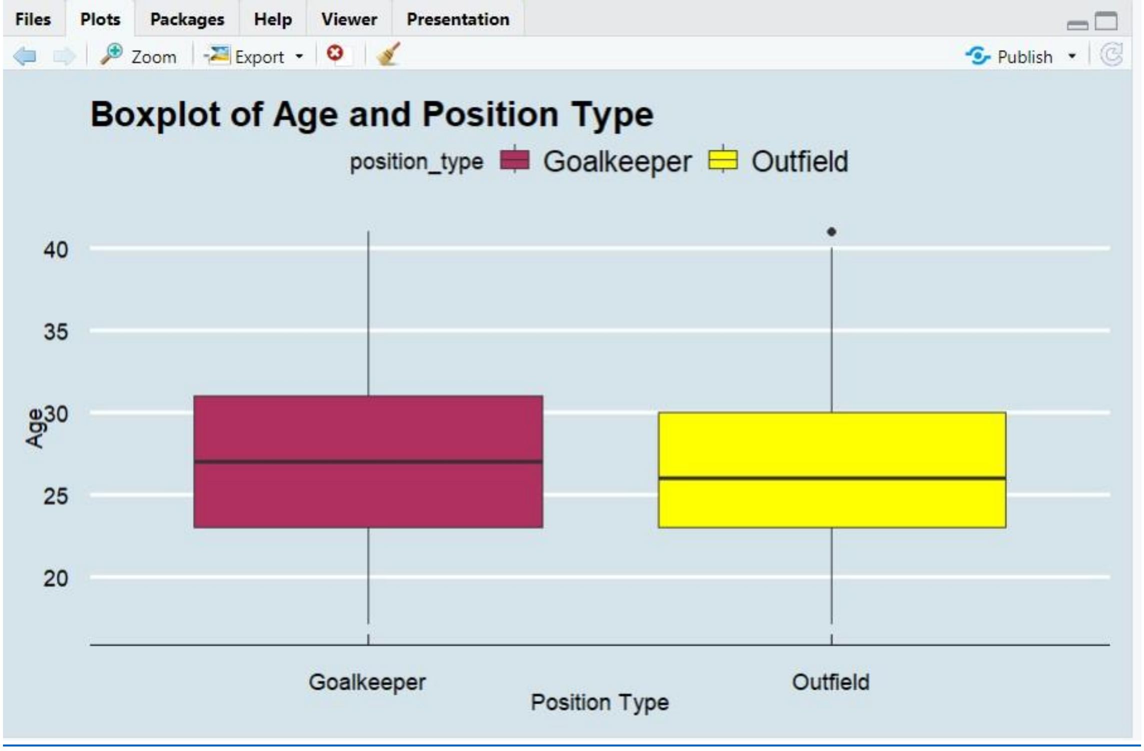
Created a similar plot for goalkeepers with their Positioning and Reflexes, and the findings were comparable.



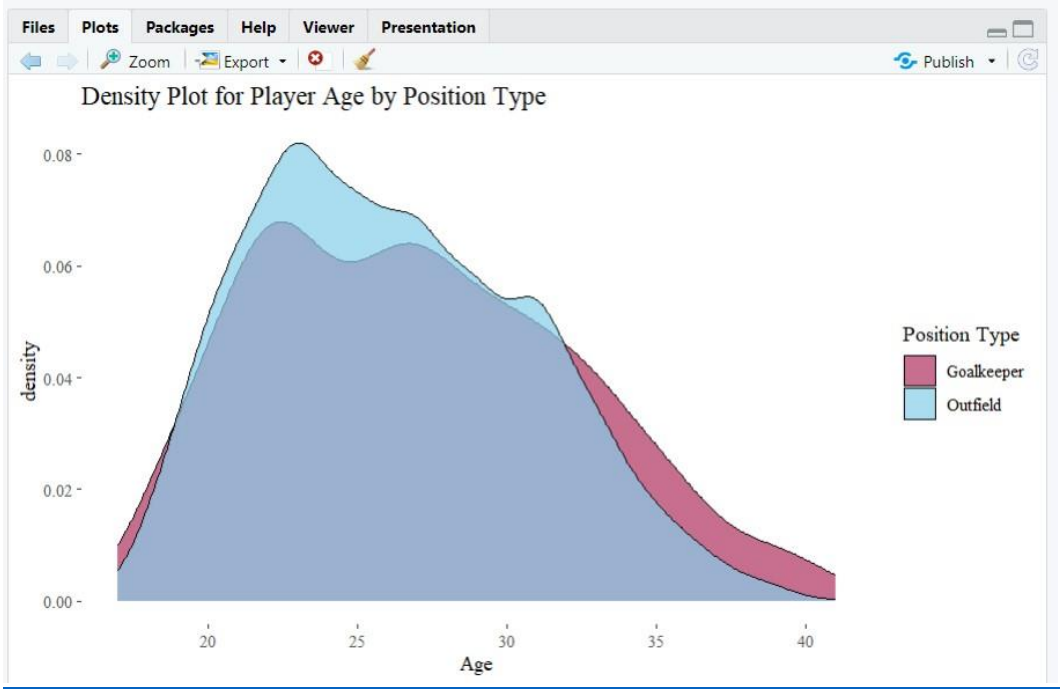
For each position type, we created boxplots of player heights and weights. The results showed that goalkeepers generally have higher quartiles, median weights, and heights compared to outfield players. This indicates that goalkeepers tend to be taller and heavier than outfielders. The height difference is particularly significant, with most goalkeepers being notably taller than outfield.




Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Boxplot of Weight and Position Type
position_type
Goalkeeper
Outfield
100
90
80
70
60
Weight
Goalkeeper
Outfield
Position Type


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Boxplot of Age and Position Type
position_type
Goalkeeper
Outfield
40
35
30
25
20
Age
Goalkeeper
Outfield
Position Type

With that created a density plot that displays the age distribution of football players categorized by their position type (goalkeeper and outfield). The plot reveals that outfield players (blue) are predominantly younger, with a peak around ages 22 to 27. In contrast, goalkeepers (red) show a wider age distribution, peaking slightly later.



Summary

By this analysis, we successfully cleaned the data and gained a thorough understanding of the various variables in the dataset. We created new columns that enhanced our analysis and generated descriptive statistics for the variables in our data. By visualizing the data, we uncovered several interesting insights, such as the age distribution of the players, as well as the height and weight distribution. Additionally, we explored the nationalities of the highest and lowest valued players and identified the total number of players from major footballing countries. We also observed a positive relationship between ball control and dribbling attributes in players, as well as between goalkeeper positioning and reflexes. Moving forward, I plan to investigate the relationships between player traits further and how these traits impact player value.

Citations

R Documentation, An introduction to R. Retrieved 01st June 2024 from https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Related-software-and-documentation

Dataset Reference, FIFA 24 Players Dataset. Retrieved 01st June 2024 from https://www.kaggle.com/datasets/rehandl23/fifa-24-player-stats-dataset/data