Football Player Market: EDA

Firass Elhouat, Marco Palmisano, Brunett Lex

1. Introduction

The main purpose of this EDA is to explore the **Football Data from Transfermarkt which we had sourced from Kaggle. The various data sets found on the site, consisted of 10 data sets, all of which provided us information regarding players'** performance metrics such as their play time during a match, goals scored, number of assist's, and any fouls incurred during a match. Other key information's provided included the team they played for, their position, foot preferences for instance whether they tend to use their right foot, left foot or both when scoring goals.

As our main purpose is to build a statistical learning model that is capable of predicting the estimated market value of a football player, it is essential to gather as much information as possible, in order to provide an accurate assessment of how these player's are valued on the football market, and to outline key variables associated with the market value's of these players. Understanding these relationships can provide insights into how different factors influence a player's market value.

2. Football Data

The dataset we decided to use, consists of 27,754 observations and 23 variables, this is after cleaning the data and merging it with other **Football Data from Transfermarkt.** The each observation is a player with the various performance metrics, team they play for, their age, and other information which we will touch up on in the next part, in explaining what each variable means.

```
# Importing data from local device
FinalData <- read_excel("~/Documents/Loyola /Fall Semster/Introduction to Predictive Analy
head(FinalData)
```

```
# A tibble: 6 x 23
  player_id player_code  citizenship birth_year last_season   age position foot
```

```
      <dbl> <chr>        <chr>            <dbl>       <dbl> <dbl> <chr>    <chr>
1        10 miroslav-kl~ Germany           1978        2015    37 Attack   right
2        26 roman-weide~ Germany           1980        2017    37 Goalkee~ left
3        80 tom-starke   Germany           1981        2017    36 Goalkee~ right
4       132 tomas-rosic~ Czech Repu~       1980        2015    35 Midfield both
5       162 marc-ziegler Germany           1976        2012    36 Goalkee~ right
6       215 roque-santa~ Paraguay          1981        2015    34 Attack   right
# i 15 more variables: height_in_cm <dbl>,
#   current_club_domestic_competition_id <chr>, Goals <dbl>, Yellow <dbl>,
#   Assist <dbl>, Minutes <dbl>, GP <dbl>, avgGoalsGame <dbl>,
#   avgMinperGoal <dbl>, avgMinPerYellow <dbl>, avgAssistmin <dbl>,
#   avgAssistGame <dbl>, avgMinutes <dbl>, market_value_in_eur <dbl>,
#   highest_market_value_in_eur <dbl>
```

2.1. Variables in the Dataset

- player_id: A unique identification number associated a player.
- player_code: The name associated to a player
- citizenship: The citizenship of the player, and the country that they represent
- birth_year: The year the player was born
- last_season: The last season the player participated, this provides insight on active and inactive players.
- age: Current age of the player, this can provide us with insight on how age impacts the player market value
- position: The player's playing position, this includes if a plays as a goalkeeper, midfield, defender or striker.
- foot: Indicates to us the player's dominant foot, ranges from left, right or both.
- height_in_cm: The player's current height in cm, a player's height may influence their value, especially in different positions such as goalkeeper, or striker.
- current_club_domestic_competition_id: A unique identification number associated by the league that the player is part of.
- Goals: Total number of recorded goals scored by a player throughout the course of their professional career.
- Yellow: Total number of yellow cards received by a player. A yellow card is received when a player committed a foul that the referee would give out a yellow card for.

- Assist: Total number of recorded assist's received by a player. This occurs when a player directly contributed to a goal scored by another player. For instance, passing the ball to a player who then scores.
- Minutes: Total number of minutes played throughout the course of their professional career.
- GP: Total number of games played throughout the course of their professional career.
- avgGoalsGame: Average number of goals scored per game
- avgMinperGoal: Average number of goals per minute per game
- avgMinPerYellow: Average number of yellow cards received per minute per game received
- avgAssistmin: Average number of assist's per minute
- avgAssistGame: Average number of assist's per game
- avgMinutes: Average number of minutes played
- market_value_in_eur: Current market value of a player in Euros
- highest_market_value_in_eur: The highest recorded market value in Euros of a player during their professional career.

2.2. Pivot Tables

By using pivot tables, this is a great way in summarizing the data, and highlighting certain variables that can tell us more about a player's characteristics, performance, team market value based on the player's market value.

Furthermore, this is efficient way of summarizing various variables by facilitating aggregation of the data across multiple dimensions. For instance, in the first pivot table of position_stats, we can see the goal contributions based on player positions. In this case, we can the attacking position yields the highest total goal and total assist which aligns with the expectations associated with that position. While, although goalkeepers tend to be much further away from the opposing goal, we do see 10 instances of when a goal keeper had scored, and several instances of goal assists.

The second pivot table provides a summary of player market values categorized by citizenship. With each row representing nation with the corresponding metrics indicating the total market value, average market value and the total number of players from that country. We can see that English player's leads with a total market value of **€4.87 billion**, and a average market value per player at **€3.83 million.**

The third pivot table focuses on summarizing the total and average market values of players across various domestic competitions, which are the leagues that the players are currently

playing in. This provides great insight in the demand for certain players by a specific league. For instance, GB1 leads with the highest total market value at **€13.59 billion**, as well as Average Market Value **€7.4 million**. This shows us the how certain leagues are willing to pay a great sum of money for specific players, and often could be an indicator for the level of competitiveness experienced in certain leagues.

```
# player position summarized by goals and assists
position_stats <- FinalData %>%
  group_by(position) %>%
  summarise(
    Total_Goals = sum(Goals, na.rm = TRUE),
    Total_Assists = sum(Assist, na.rm = TRUE)
  ) %>%
  arrange(desc(Total_Goals))

# Display the pivot table
print(position_stats)
```

```
# A tibble: 5 x 3
  position   Total_Goals Total_Assists
  <chr>            <dbl>         <dbl>
1 Attack           58957         32581
2 Midfield         27160         28848
3 Defender         13210         17560
4 Missing            237           136
5 Goalkeeper          10           177
```

```
# Player citizenship by average and total market values, in order to see how citizienship
market_value_citizenship <- FinalData %>%
  group_by(citizenship) %>%
  summarise(
    Total_Market_Value = sum(market_value_in_eur, na.rm = TRUE),
    Average_Market_Value = mean(market_value_in_eur, na.rm = TRUE),
    Total_Players = n()
  ) %>%
  arrange(desc(Total_Market_Value)) # Sort by total market value

# Display the pivot table
print(market_value_citizenship)
```

```
# A tibble: 180 x 4
```

```
   citizenship Total_Market_Value Average_Market_Value Total_Players
   <chr>                    <dbl>                 <dbl>         <int>
 1 England             4868900000             3833780.          1270
 2 France              4312845000             2773534.          1555
 3 Spain               4166090000             2405364.          1732
 4 Brazil              3879455000             2491622.          1557
 5 Germany             2508245000             2120241.          1183
 6 Italy               2279205000             1378829.          1653
 7 Netherlands         2237255000             1957353.          1143
 8 Portugal            2216395000             2073335.          1069
 9 Argentina           1764595000             3011254.           586
10 Belgium             1408420000             1730246.           814
# i 170 more rows
```

```r
# summarizing competition(league) by market value
competition_value_table <- FinalData %>%
  group_by(current_club_domestic_competition_id) %>%
  summarise(
    Total_Market_Value = sum(market_value_in_eur, na.rm = TRUE),
    Average_Market_Value = mean(market_value_in_eur, na.rm = TRUE),
    Total_Players = n()
  ) %>%
  arrange(desc(Total_Market_Value))  # Sort by total market value

# Display the pivot table
print(competition_value_table)
```

```
# A tibble: 14 x 4
   current_club_domestic~1 Total_Market_Value Average_Market_Value Total_Players
   <chr>                             <dbl>                 <dbl>         <int>
 1 GB1                         13588130000             7400942.          1836
 2 ES1                          6666955000             3406722.          1957
 3 IT1                          6159510000             2171135.          2837
 4 L1                           5318055000             3236795.          1643
 5 FR1                          4698590000             2379033.          1975
 6 PO1                          2537595000             1039998.          2440
 7 NL1                          1870390000             1027122.          1821
 8 TR1                          1870340000              706322.          2648
 9 BE1                          1522535000              872013.          1746
10 RU1                          1486275000              718007.          2070
11 GR1                          1050815000              463936.          2265
```

```
12 SC1                                709010000             459203.          1544
13 UKR1                               700165000             419010.          1671
14 DK1                                651405000             500696.          1301
# i abbreviated name: 1: current_club_domestic_competition_id
```

3. Data Summary

In this section we are primarily going to focus the descriptive statistics of the data, including correlations between variables, and highlighting key relationships.

3.1. Correlation plot

From the correlation matrix plot, we are mostly interested in looking if market_value_in_eur has any strong correlations with the other variables. In this, the highest being a strong correlation with highest_market_value_in_eur at a positive 0.74, this would suggest to us that players achieve a high valuation at some point in their careers tend retain this value in their career. This is better summarized when we plot the two together, which will be shown in the next few parts.

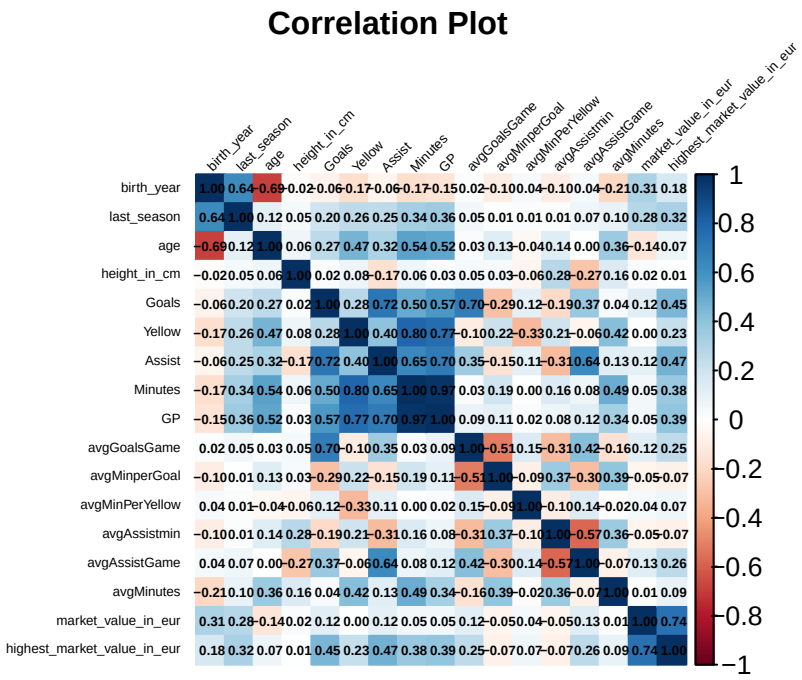
Secondly, if we would look at variables indicating player's performance such as goals, assist's, and games played, although very moderate, this generally indicates they have an association with the market value without being too strong. Furthermore, as we progress through the project, much of the variables would need to be transformed in order to capture the the true underlying relationships, and even introducing other variables that may tell us more about how player's a valued in the market.

As the dataset we are working with right, only contains a subset of variables, in this case it would require certain variables that can explain how goalkeepers and defenders are valued, as often these players tend to not score as much, however other performances metrics may aid in this discussion.

In terms of highly correlated variables, we can see a few that may need to be addressed and handled, as they may cause multicollinearity issues in our models. For instance, GP high positive correlations with a few other variables such as "Yellow" at 0.77, "Assist" at 0.70, and the highest being"Minutes" at 0.97. If the VIF level exceeds a certain threshold for instance between 5 to 10, this could inflate our standard errors, nonetheless, proper handling of such variables will need to be highlighted during modeling, in order to produce an accurate model.

```r
# extracting just the numeric variables to compute the correlation matrix and plot
numeric_data <- FinalData[, !(names(FinalData) %in% c("player_id", "player_code", "current
numeric_data <- numeric_data[, sapply(numeric_data, is.numeric)]

cor_matrix <- cor(numeric_data, use = "complete.obs")
```

```
# plotting correlation
corrplot(cor_matrix,method = "color",addCoef.col = "black", tl.col = "black",tl.srt = 45,
```



3.2. Descriptive statistics

Looking through a short example of the variables, we can see that the age of the players are fairly symmetric in distribution, while the age ranging from 15 to 44 is somewhat showing variability.

In terms of height_in_cm, there is an error most probably an input mistake, in which the min is 17 cm which is highly unlikely to be accurate. Overall, fairly symmetric in distribution, with the max height at 207 cm, and the median height of players being 182 cm.

Looking at the market value in euros, the mean is significantly higher then the median, this could indicate a right-skewed distribution, with the highest paid players reaching 200,000,000 euros.

```
summary(FinalData$age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   21.00   25.00   25.44   29.00   44.00
```

```r
summary(FinalData$height_in_cm)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   17.0   178.0   182.0   182.3   187.0   207.0
```

```r
summary(FinalData$market_value_in_eur)
```

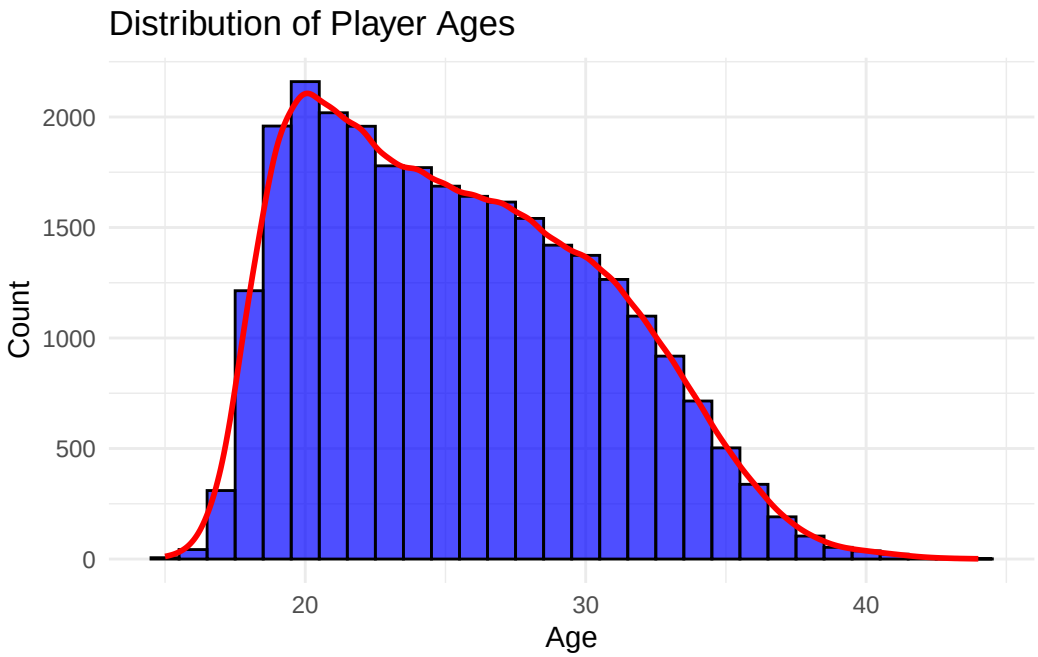
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    10000    100000    250000   1759378    750000 200000000
```

Data Visualization & Exploration

```r
# Histogram of players by their current age
ggplot(FinalData, aes(x = age)) +
  geom_histogram(aes(y = ..count..), binwidth = 1, fill = "blue", color = "black", alpha =
  geom_density(aes(y = ..count.. * (binwidth = 1)), color = "red", size = 1, adjust = 1) +
  labs(title = "Distribution of Player Ages", x = "Age", y = "Count") +
  theme_minimal()
```

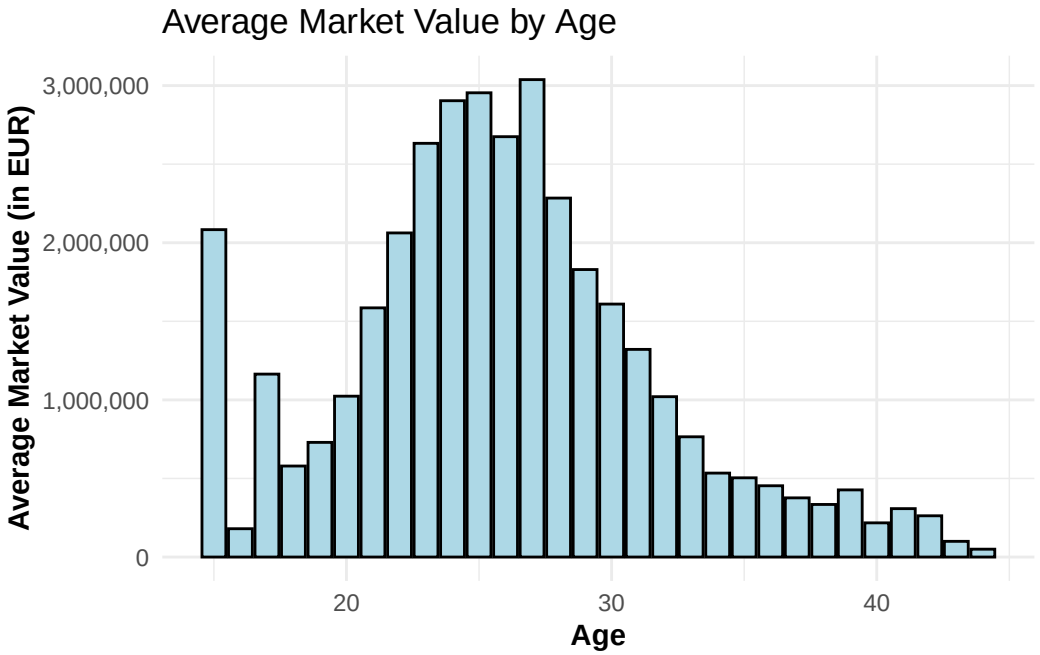
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(count)` instead.
```



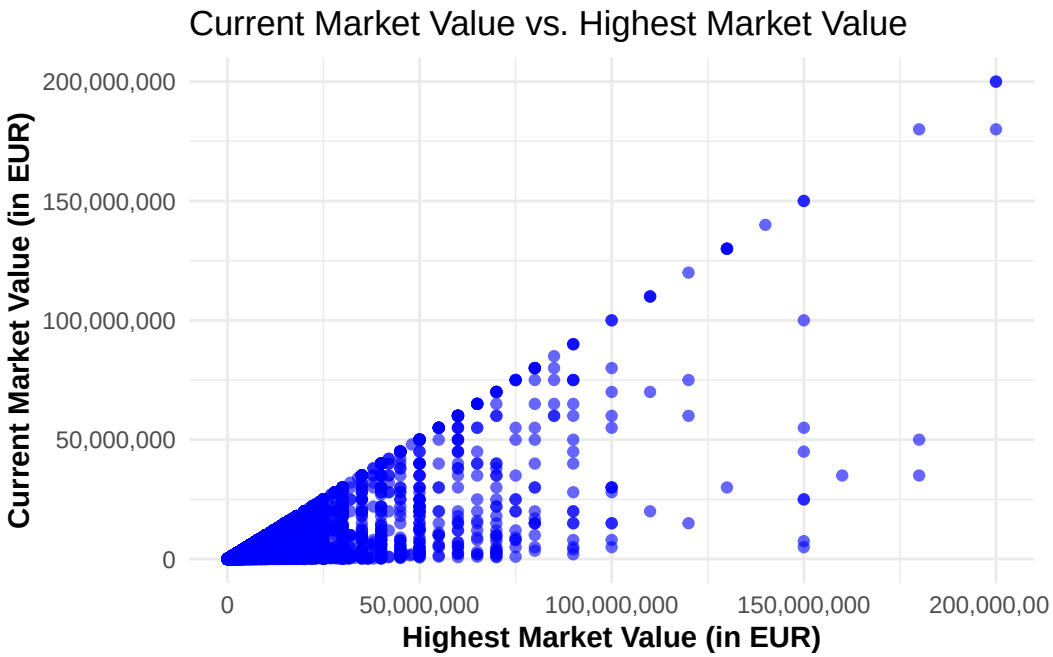
The plot above we can see how the age's of players is right-skewed, with a high concentration of players in their early twenties and fewer players in their late thirties, which a fairly common thing in most sports.

```r
# Bar Plot of Current Market Value by Age
FinalData %>%
  group_by(age) %>%
  summarise(avg_market_value = mean(market_value_in_eur, na.rm = TRUE)) %>%
  ggplot(aes(x = age, y = avg_market_value)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "black") +  # Bar plot
  labs(
    title = "Average Market Value by Age",
    x = "Age",
    y = "Average Market Value (in EUR)"
  ) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(face = "bold"),  # Bold x-axis title
    axis.title.y = element_text(face = "bold")
  ) +
  scale_y_continuous(labels = scales::comma)
```



In the plot above, this is great illustration in how the an almost bell shaped distribution of player's age by their market value, with the player's value peaking at the end of their twenties, and followed by a decline, which is fairly logical as player's age their performance on the pitch is expected to decline, which will ultimately reduce their market demand.

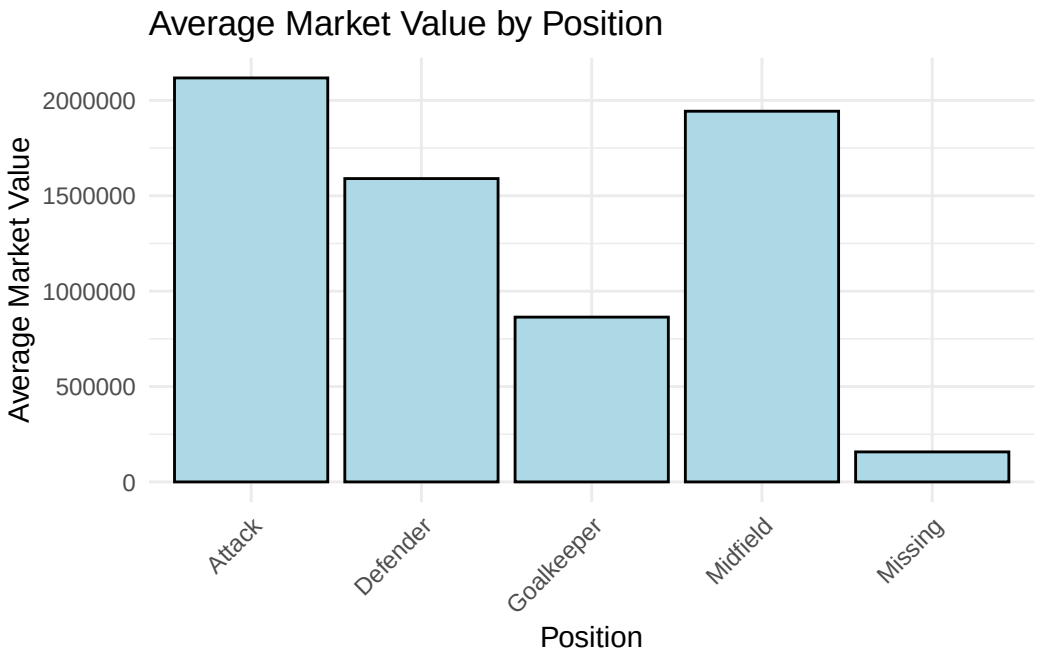
```
# Bar Plot of Current Market Value by Age
ggplot(FinalData, aes(x = highest_market_value_in_eur, y = market_value_in_eur)) +
  geom_point(alpha = 0.6, color = "blue") +  # Adjust transparency and color
  labs(
    title = "Current Market Value vs. Highest Market Value",
    x = "Highest Market Value (in EUR)",
    y = "Current Market Value (in EUR)"
  ) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold")
  ) +
  scale_x_continuous(labels = scales::comma) +  # Format x-axis labels with commas
  scale_y_continuous(labels = scales::comma)
```



We had discussed earlier about the strong strong correlation that market_value_in_eur and highest_market_value_in_eur had a positive 0.74. In the plot above, this show us the upwards trend that often when players reach a peak value, they are able to maintain this value over time.

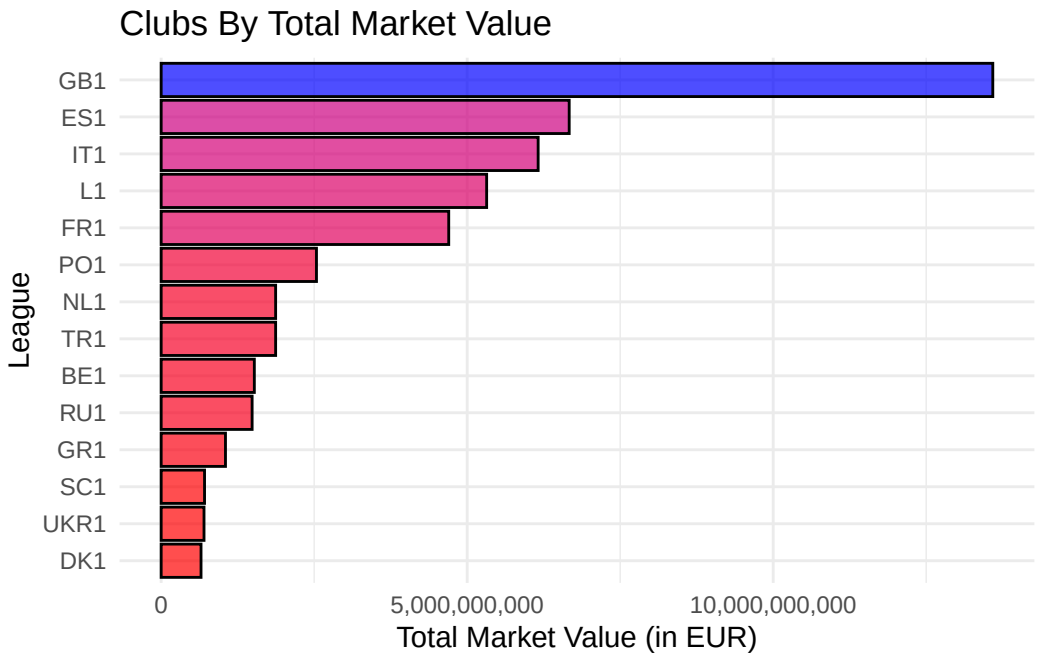
In the plot above, we can clearly observe this upward trend, indicating that as players reach their peak valuations, their current market values tend to remain elevated.

```
# computing average current market value by player position
position_summary <- FinalData %>%
  group_by(position) %>%
  summarise(avg_market_value = mean(market_value_in_eur, na.rm = TRUE))

# barplot of average market value by player position
ggplot(position_summary, aes(x = position, y = avg_market_value)) +
  geom_bar(stat = "identity",fill = "lightblue", color = "black") +
  labs(title = "Average Market Value by Position", x = "Position", y = "Average Market Val
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



In the plot above, we assess how a player's position correlates with their average market value. Notably, the "Attack" position yields the highest average market value, exceeding €2 million Euros, followed closely by the "Midfield" position, which approaches this figure. This trend highlights the significant financial investment clubs make in attacking and midfield players, as they are crucial for creating scoring opportunities and contributing to the team's overall performance.

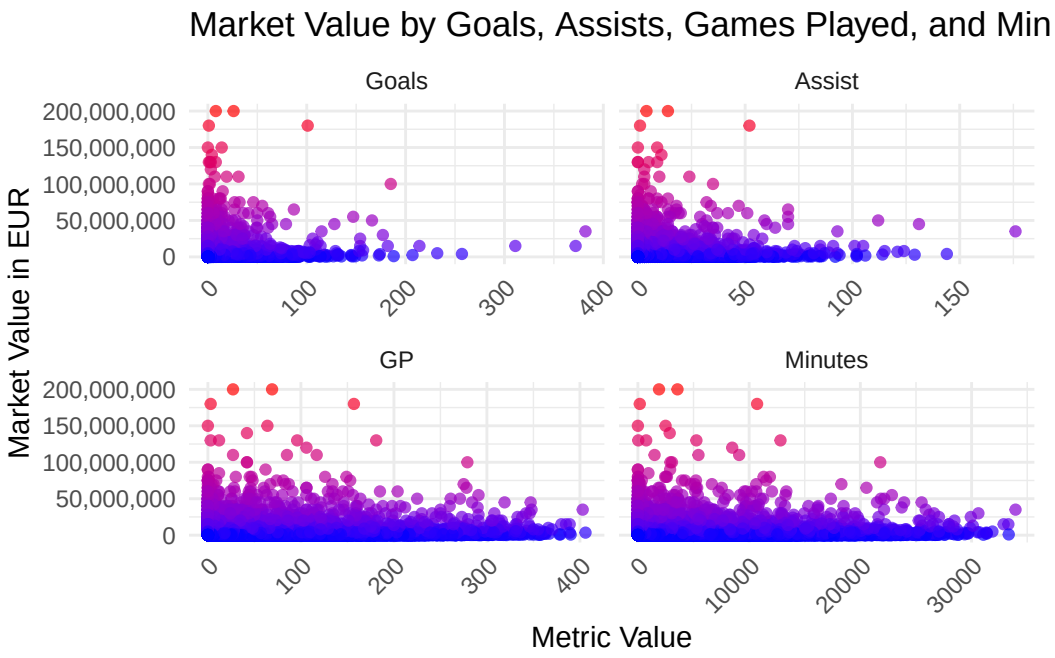
```r
# summarizing top 10 leagues by market value of the players within those leagues
TopLeagues <- FinalData %>%
  group_by(current_club_domestic_competition_id) %>%
  summarise(total_value = sum(market_value_in_eur, na.rm = TRUE)) %>%
  arrange(desc(total_value))

ggplot(TopLeagues, aes(x = reorder(current_club_domestic_competition_id, total_value), y =
  geom_bar(stat = "identity", color = "black", alpha = 0.7) +  # Remove fill color here
  coord_flip() +
  labs(title = "Clubs By Total Market Value", x = "League", y = "Total Market Value (in EU
  theme_minimal() +
  scale_y_continuous(labels = scales::comma) +
  scale_fill_gradient(low = "red", high = "blue") +
  theme(legend.position = "none")
```



In terms of examining if certain leagues effect a player's market value, we can see from the plot above, that League GB1 is significantly exceeds the other leagues in terms of market value of the players who play in that league. This indicates to us that player's in this league are highly demand, particularly in terms of market demand, and competitiveness experience in the league.

```
# plot data
plot_data <- FinalData[, c("market_value_in_eur", "Goals", "Assist", "GP", "Minutes")]

# Melt data into long format for easier plotting
plot_data_long <- melt(plot_data, id.vars = "market_value_in_eur")

# Plot each predictor on the x-axis with market value on the y-axis
ggplot(plot_data_long, aes(x = value, y = market_value_in_eur, color = market_value_in_eur
  geom_point(alpha = 0.7) +  # Use color gradient for points
  facet_wrap(~ variable, scales = "free_x") +
  labs(
    title = "Market Value by Goals, Assists, Games Played, and Minutes",
    x = "Metric Value",
    y = "Market Value in EUR"
  ) +
  scale_y_continuous(labels = comma) +  # Format y-axis with commas
```

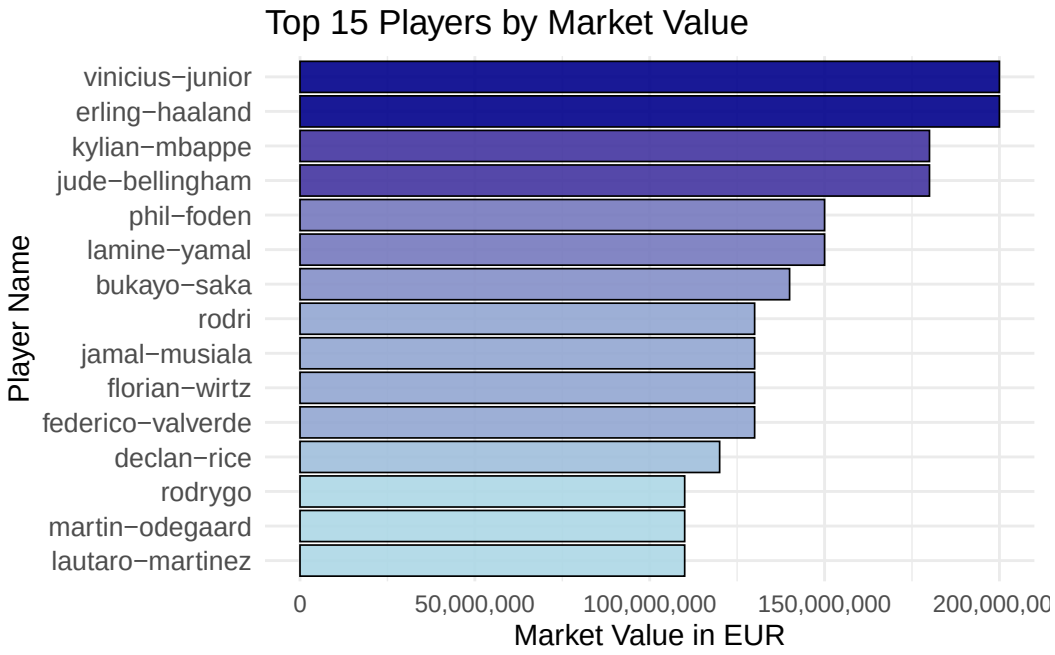
```
  scale_color_gradient(low = "blue", high = "red") +  # Custom color gradient
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  theme(legend.position = "none")
```



In this plot above, we wanted to asses how player's metrics may show how it impacts their Market value, we can see that although certain players had scored close to 400 goals during their seasons, their current market value is has not increased. This could tell us that these are probably players who have reached a certain age in which may impact their value in the market. Certain transformation and feature engineering techniques would be needed in order to further help the model explain the market value, for instance excluding players who had stopped playing for a certain period, or had completely retired.

```
# summarizing the top 15 players by market value
top_players <- FinalData %>%
  arrange(desc(market_value_in_eur)) %>%
  head(15)

ggplot(top_players, aes(x = reorder(player_code, market_value_in_eur), y = market_value_in
  geom_bar(stat = "identity", color = "black", size = 0.3, alpha = 0.9, aes(fill = market_
  coord_flip() +  # Flip coordinates for better readability
```

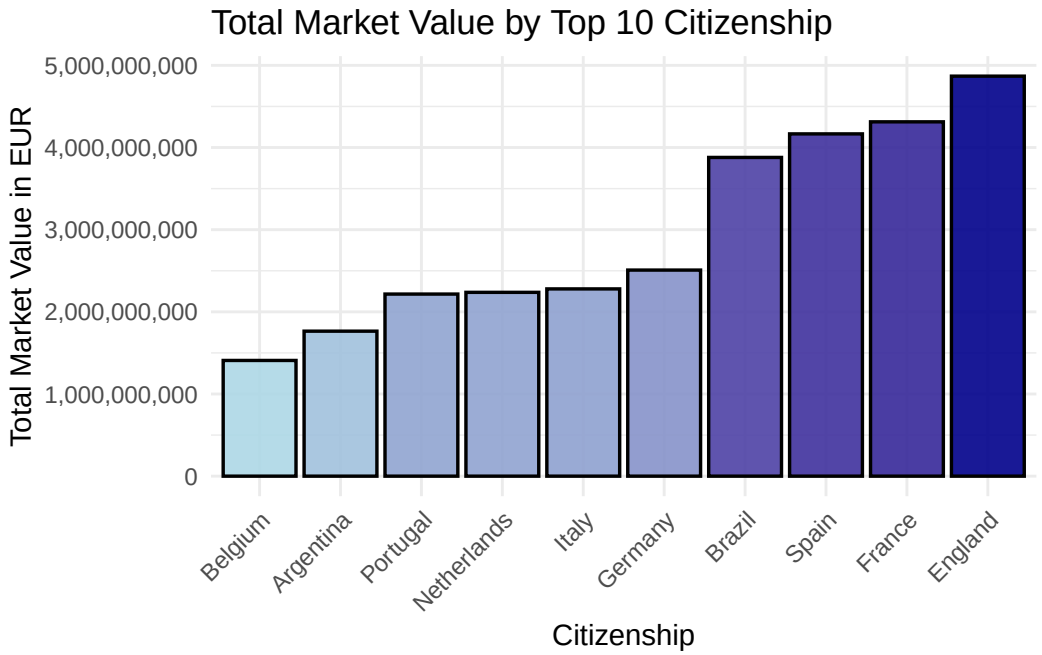
```
  scale_y_continuous(labels = scales::comma) +  # Format y-axis labels with commas
  labs(x = "Player Name", y = "Market Value in EUR", title = "Top 15 Players by Market Val
  scale_fill_gradient(low = "lightblue", high = "darkblue") +  # Custom color palette
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10)) +
  theme(legend.position = "none")
```



In the bar plot comparison, highlighting players like **Vinicius Junior** and **Erling Haaland** at the maximum market value of €200 million emphasizes the very top end of the market value distribution. Both players represent a rarity in the dataset due to their exceptionally high valuation, which distinctly skews the mean market value upwards.

```
top_citizenship <- FinalData %>%
  group_by(citizenship) %>%
  summarise(total_market_value = sum(market_value_in_eur, na.rm = TRUE)) %>%
  arrange(desc(total_market_value)) %>%
  slice_head(n = 10)

# Plotting Total Market Value by Top 15 Citizen ZIP
ggplot(top_citizenship, aes(x = reorder(citizenship, total_market_value), y = total_market
  geom_bar(stat = "identity", color = "black", size = 0.6, alpha = 0.9) +  # Use fill aest
```

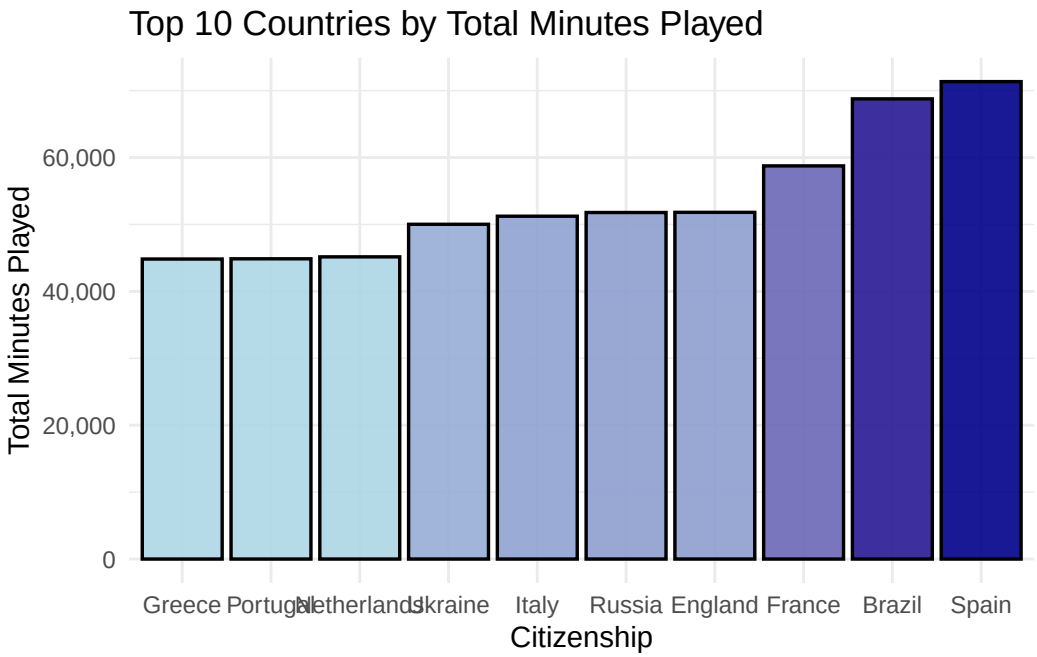
```
  labs(x = "Citizenship", y = "Total Market Value in EUR", title = "Total Market Value by
  scale_y_continuous(labels = comma) +  # Format y-axis labels with commas
  scale_fill_gradient(low = "lightblue", high = "darkblue") +  # Custom color palette for
  theme_minimal() +  # Use a minimal theme for a cleaner look
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(legend.position = "none")
```



From this bar plot, we wanted to see the total market valued by citizenship, and from the plot we can see that English players achieved the highest sum of the market value almost reaching €5 Billion Euros, this is often the case seen that English players tend to go for a higher prices in the football markets in comparison to player's coming in from Belgium.

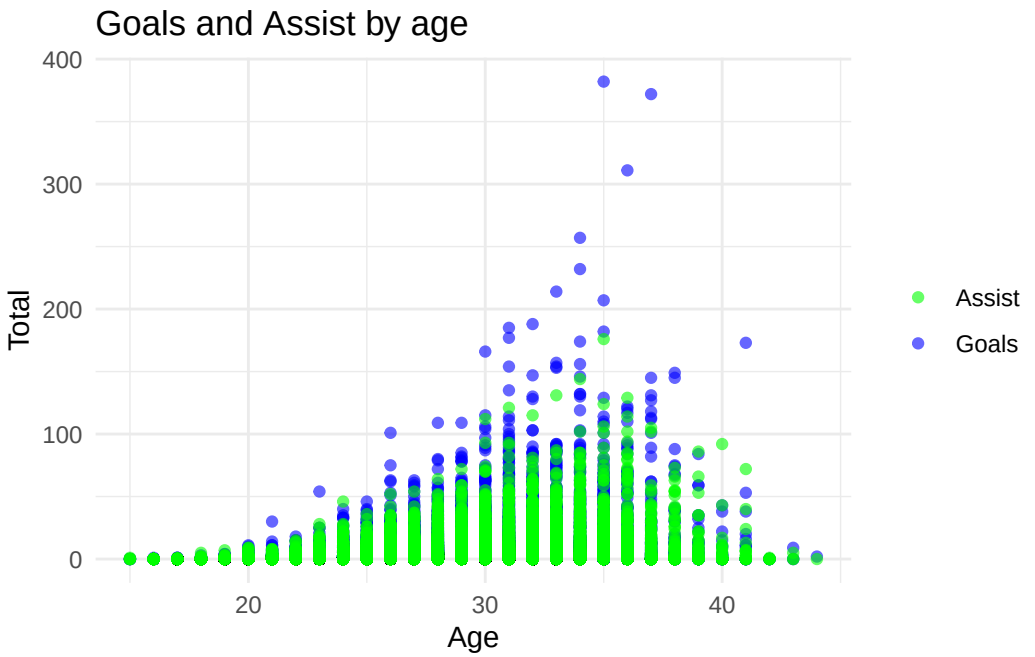
```
# finding the top 10 Citizenship by average minutes played
FinalData %>%
  group_by(citizenship) %>%
  summarise(total_minutes = sum(avgMinutes, na.rm = TRUE)) %>%
  top_n(10, total_minutes) %>%
  ggplot(aes(x = reorder(citizenship, total_minutes), y = total_minutes, fill = total_minu
  geom_bar(stat = "identity", color = "black", size = 0.6, alpha = 0.9) +
  labs(x = "Citizenship", y = "Total Minutes Played", title = "Top 10 Countries by Total M
  scale_y_continuous(labels = scales::comma) +
```

```
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  theme_minimal() +
  theme(legend.position = "none")
```



From the graph above, we can see the comparison that although Spanish players playing the most, they still lag behind English and French players in terms of total market value. This would suggest, that although Spanish players are frequently present in games, their average market value may not always be as high as those of English or French players.

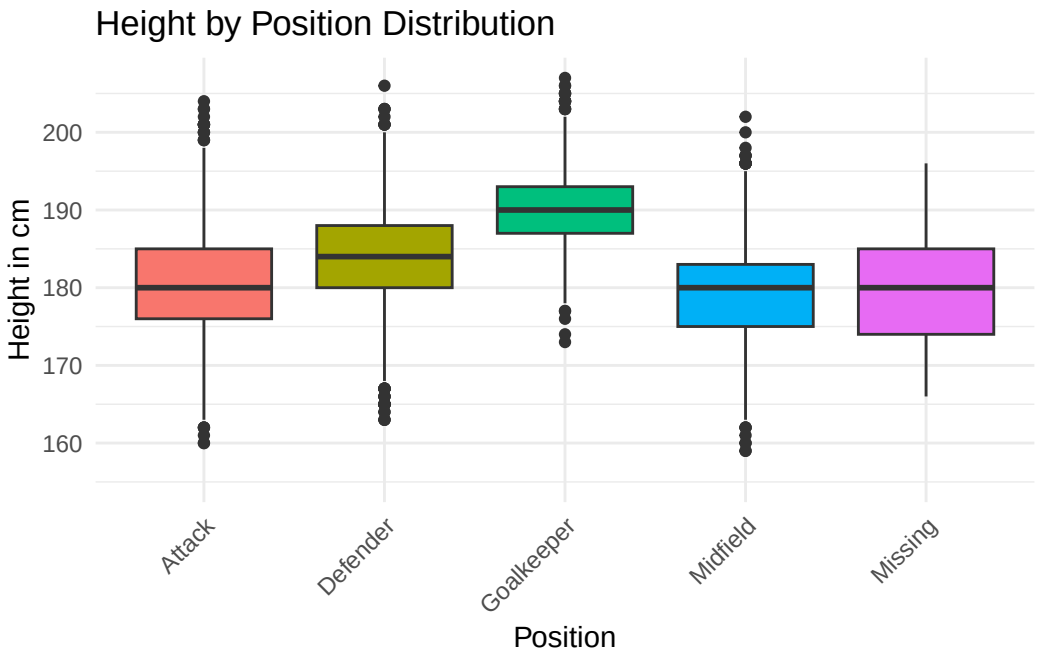
```
#Relation of Goals/Assist by age
ggplot(FinalData, aes(x = age)) +
  geom_point(aes(y = Goals, color = "Goals"), alpha = 0.6) +
  geom_point(aes(y = Assist, color = "Assist"), alpha = 0.6) +
  labs(title = "Goals and Assist by age", x = "Age", y = "Total") +
  scale_color_manual(values = c("Goals" = "blue", "Assist" = "green")) +
  theme_minimal() +
  theme(legend.title = element_blank())
```



According to the graph, players over 30 years old and less than 40, show a better performance when it comes to assists and scores, therefore a player between this age bracket has a higher market value than younger players

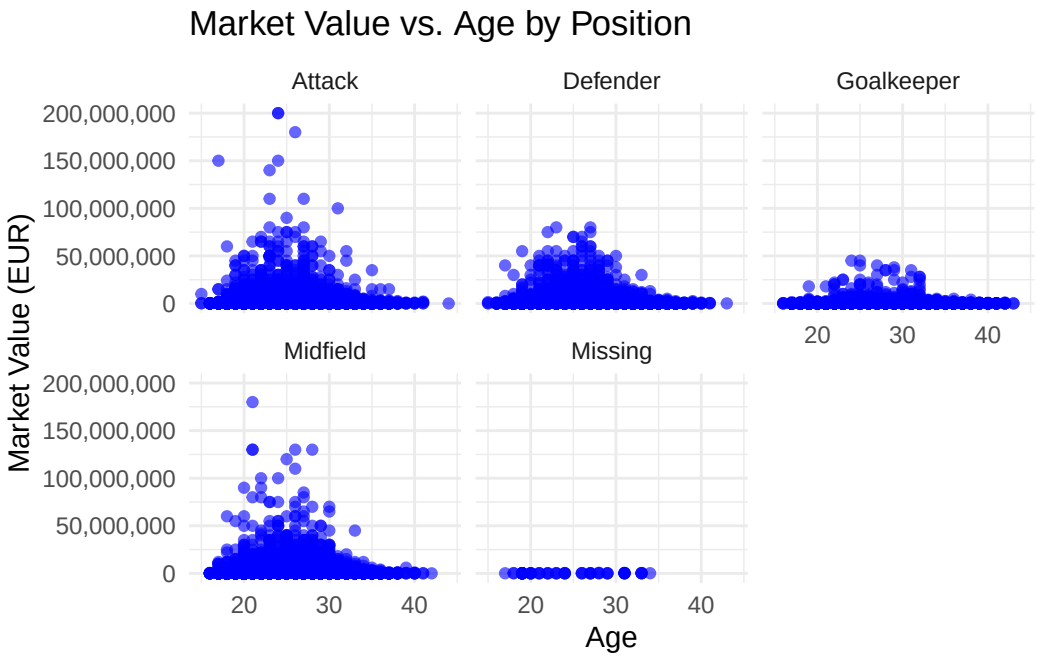
```
ggplot(FinalData, aes(x = position, y = height_in_cm, fill = position)) +
  geom_boxplot() +
  labs(title = "Height by Position Distribution", x = "Position", y = "Height in cm") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(legend.position = "none") +
  scale_y_continuous(limits = c(155
                                , NA))
```

```
Warning: Removed 4 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



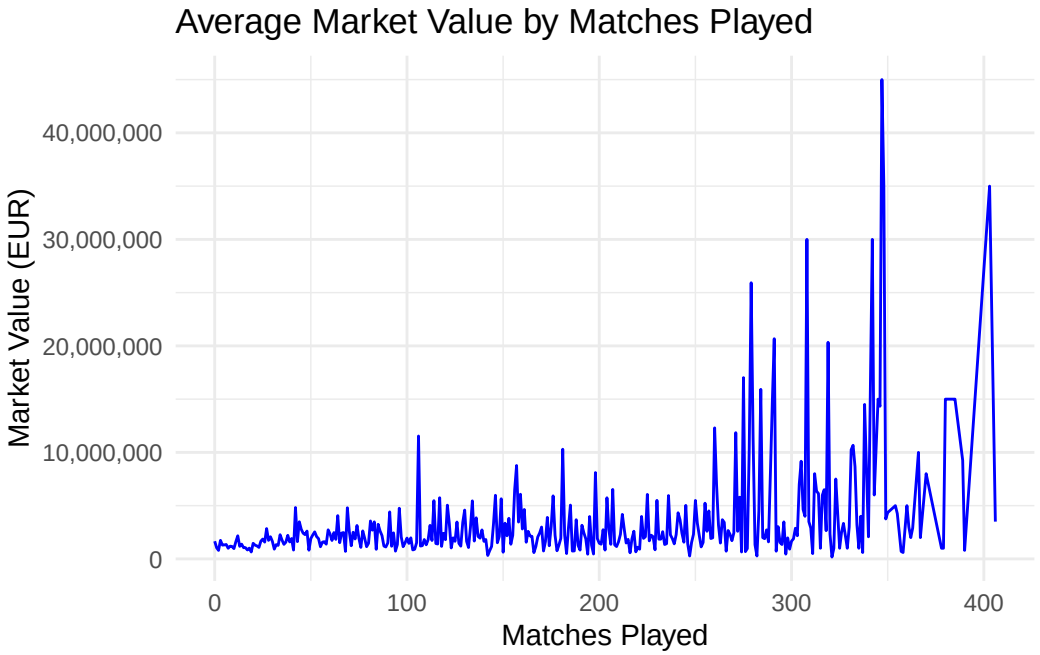
According to the boxes, Goalkeepers are the tallest players with a media height near to 190cm and a few over 2 meters. Defenders comes second with a approximate media height of 185 and a minimum of 180 for 75% of cases, in third place we found attack players, with at least 50% of this group with a height of 180 cm. Finally the midfield positions shows the shortest players with more than 50% of this group with less than 180 cm of height.

```
#Market Value vs age by position
ggplot(FinalData, aes(x = age, y = market_value_in_eur)) +
  geom_point(color = "blue", alpha = 0.6) +
  facet_wrap(~position) +
  labs(title = "Market Value vs. Age by Position", x = "Age", y = "Market Value (EUR)") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```



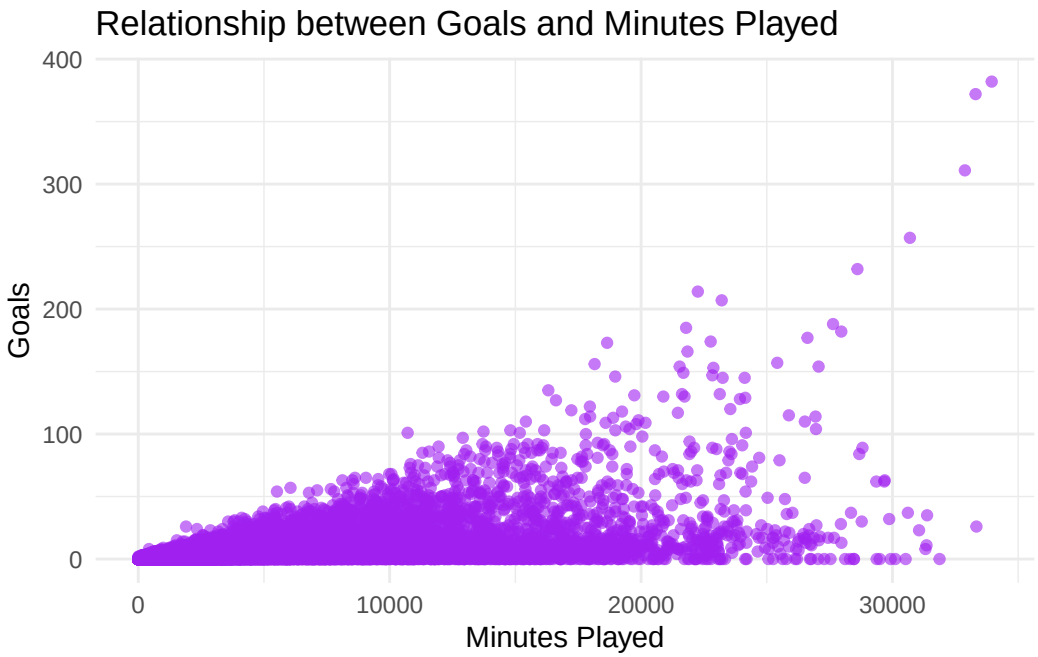
The graphs suggest that the players with highest market value according to the role they play are the 'Attack' players, specially the ones between 20 and 30 years old, reaching as far as 200 million euros per player, then the 'Midfield' players come as second in market value; the third place gods for defenders and finally the players with lowest market value are de goalkeepers. For all positions is important to highlight that the age bracket between 20's and 30's stay consistent as the most valuable players

```
# Average Market Value by Matches Played
ggplot(FinalData, aes(x = GP, y = market_value_in_eur)) +
  stat_summary(fun = mean, geom = "line", color = "blue") +
  labs(title = "Average Market Value by Matches Played", x = "Matches Played", y = "Market
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```



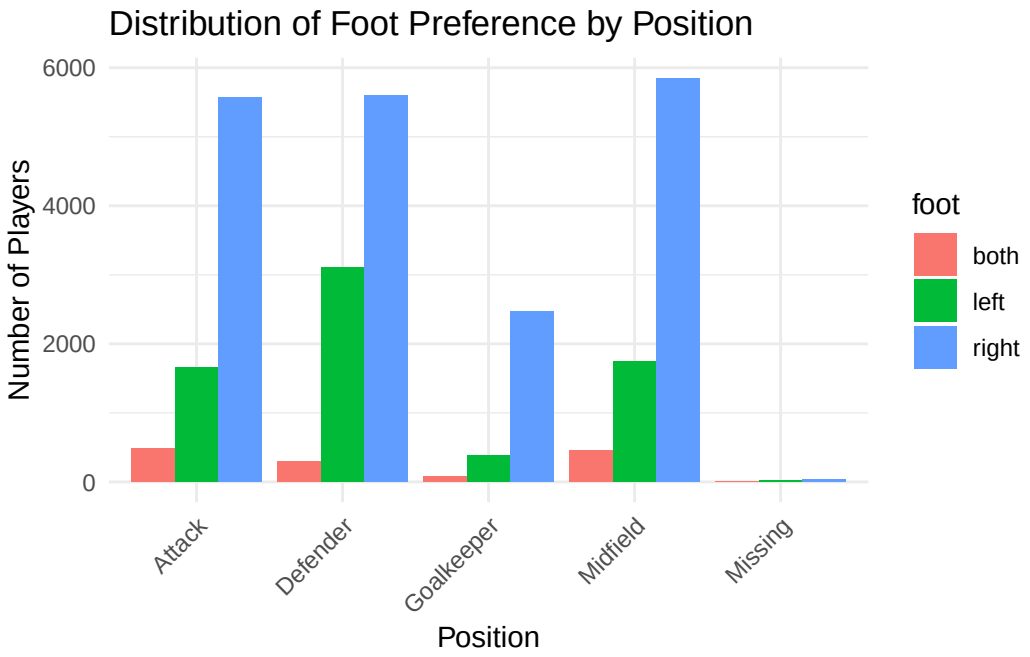
The Scale shows that there is a tendency to increase the market value for each player according to the number of matches played. However, this tendency reach the highest point at approximately 350 matches, and starts decreasing as the players hit the 400 matches, this way, we can conclude that the experience is directly proportional with the players market value until they go over 350 matches. Analyzing the dynamics showcased in the previous graphs we can conclude that the decrease in market value is because as they past 350 matches played, players reach their 40's and decrease their market value and scoring.

```
ggplot(FinalData, aes(x = Minutes, y = Goals)) +
  geom_point(alpha = 0.6, color = "purple") +
  labs(title = "Relationship between Goals and Minutes Played", x = "Minutes Played", y =
  theme_minimal()
```



According to the plot, the chance of scoring is proportional to the amount of minutes played, nevertheless we see a clear dispersion as players reach higher scores records.

```
# Foot Preference by Position
ggplot(FinalData, aes(x = position, fill = foot)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of Foot Preference by Position", x = "Position", y = "Number
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The distribution shows that for all positions, most players prefer their right foot over left, or both. For the group of players that prefer their left foot, Defenders show the highest preference among the rest of the positions, followed by Midfield and Attack players. Finally, there is a small group of players in all positions that has no preference as they play with both feet.

4. Model Approaches and Feature Engineering

The exploration of the player market values has a revealed a few key insights in how the structure of the relationships among the various variables, such as their ages, league, position and other performance players metrics. To better capture and leverage these relationships, we would need assess them through several model approaches, while also engaging in feature engineering, in terms of which variables may need to put aside, and derive new ones, while introducing the other subset variables which have not been included yet.

4.1 Model Approaches

In order to capture the true underlying relationship, we propose using several types of statistical learning appraoches without confining ourselves to a single approach. For instance, we could use a multiple linear regression, in case the relationship is linear. However, a few predictors have shown a different type, we brings us to possibility of using a non-linear approach such as Random Forest.

By using Random Forest, it is capable of handling the non-linearities and interactions among the predictors more effectively, or even using SVM's (Support Vector Machines). Overall, by listing the number of approaches that can be made, it opens up several ways of handling the complexity of this dataset, and even so, it would require excluding or including certain variables that may contribute to the model.

4.2 Feature Engineering

In most cases, certain variables do need to be derived or transformed in order for the model to better handle it. This includes feature scaling techniques such as standardization, and min-max scaling. While in this case, we would want to include variables that captures the stage of a player's career such as when did they debut, peak performance years, injuries, and other player performance metrics such as expected goals. Furthermore, as we had noticed much of the metrics are complimenting player's in the "attack" position, we would need to include variables that are better suited for other position's as well such as goalkeeper's performance metrics, this will further help the model make accurate predictions in the market value.