Project Proposal

Paulo Souza

Dataset Information

The dataset contains information about the players that have won the NBA's Men of the Week's award. The NBA (National Basketball Association) is a men's professional basketball league in North America; composed of 30 teams (29 in the United States and 1 in Canada). It is widely considered to be the premier men's professional basketball league in the world [Wikipedia]. This dataset was obtained from Kaggle.

Dataset Semantics

Each row of the dataset contains the following information about each player awarded with the Player of the Week prize:

- Age (player age at the time)
- Conference (East/West/NaN)
- Date (award date)
- Draft Year
- Height (in feets)
- Player
- Position
- Season
- Season short (season ending year)
- Seasons in league
- Team
- Weight (in pounds)
- Real_value (If two awards given at the same week [East & West] the player got 0.5, else 1 point)

Data Analysis

```r
rm(list=ls())
library(dplyr)
library(ggplot2)
```

Firstly, I read the dataset and call the `summary()` function in order to get some insights about the data.

According to the `summary()` function output, there are 384 entries in the dataset in which the Conference field is not filled. Hence, I'm gonna print the first lines of the dataset with the `head()` function to check how I could fix it.

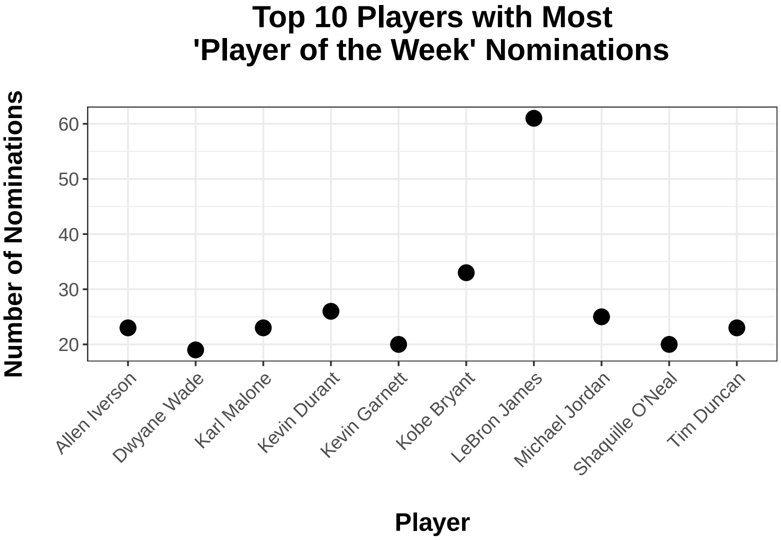
The output of the `head()` function was actually very useful since I realized that I could just use the team name as a reference in order to fill the empty Conference fields. I mean, as far as I know, most of the teams active nowadays already exist back in the 80s (date of the first records in the dataset), so I could just compare the teams of the filled rows with the teams of the empty ones, and if the team is the same in both lines, then I should be able to get the name of the conference of those empty rows. In order to do that, I'm gonna create a loop that goes through each row of the dataset that contains a value in the Conference field, then I just have to tell R to replace the field Conference of each other lines of the same team with the value I get from the Conference field of that line.

Even doing the aforementioned adjust, I realized that there are still 8 entries without a value in the Conference field. Thus, I decided to print those values in order to check what I could do.

According to the output of the previous command, only the entries of Washington Bullets players are still without content in the Conference field. That's a little awkward because I don't believe none of the players of this team own the Player of the Week since 1997. Then, I decided to do some research on Washington Bullets: According to Wikipedia, in 1997 the director board of the Washington Bullets rebranded themselves as the Wizards.

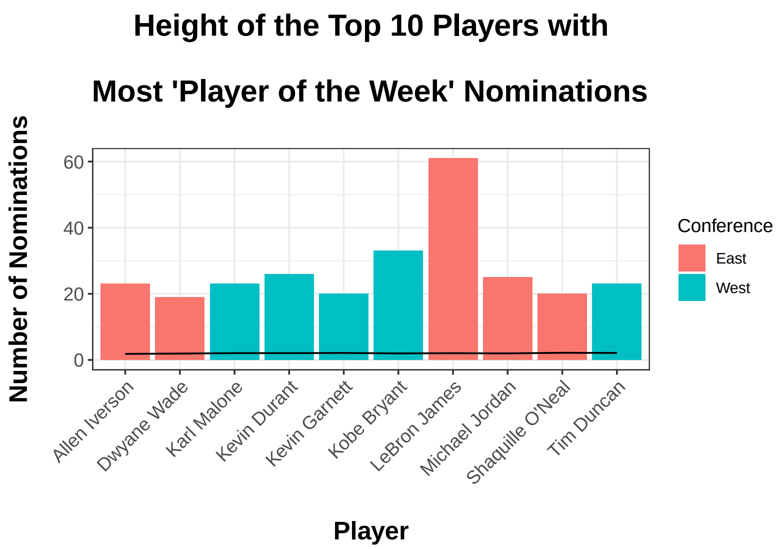
Thus, I decided to print out the records of players of Washington Wizards and check which conference it belongs in order to update the Conference field in the old records of the team, and finally I use `filter()` and `count()` functions in order to check the number of entries without conference remaining.

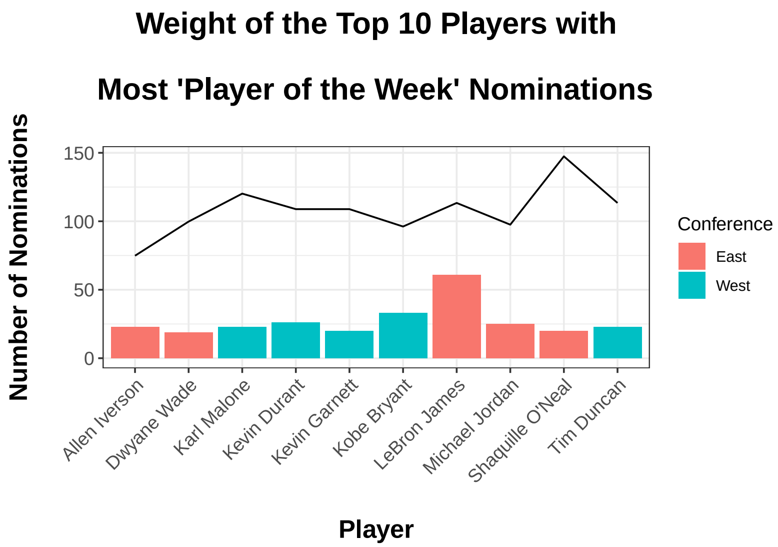
Then, I wanted to check which players most have owned the Player of the Week award.



The results were slightly surprising since Lebron James appeared with a crushing advantage above the other players. After doing some research on the possible reasons for that result, I discovered that Lebron James' team (Cleveland Cavaliers) won the last 5 titles of his conference. Before joining the Cavaliers, Lebron played for Miami Heat (from 2010-2014), and that team won its conference league 4 times in this period. Since the team performance directly affects the nomination to Player of The Week, it is reasonable that Lebron James to lead the ranking.

Moreover, I decided to analyze which characteristics are responsible for building that huge difference. When we are talking about people characteristics, one of the first measures that come into my mind is height and weight. Then, I decided to create charts to analyze the correlation between height and weight with the number of nominations to Player of The Week. In addition, I created functions to format height to meters and weight to kilos.



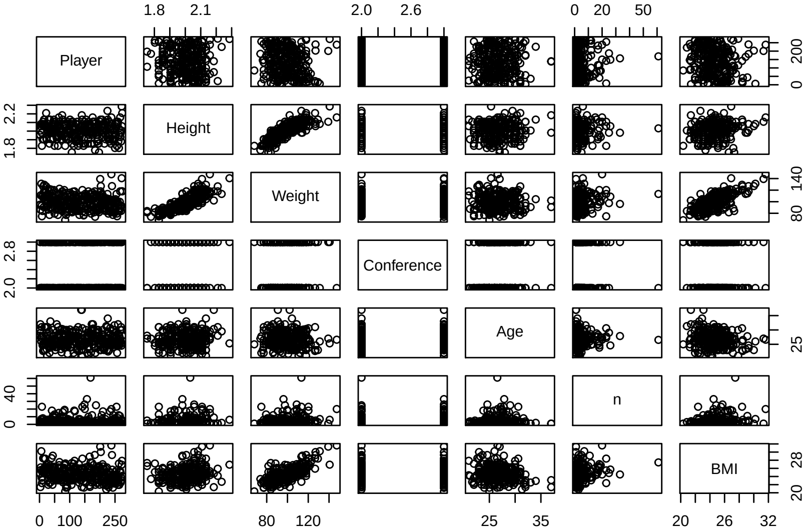


In addition, I wanted to visualize the difference of these players regarding phisical attributes in a more specific way, so I decided to calculate the BMI (Body Mass Index), which is given by weight (in kilos) multiplied by height (in meters) squared:

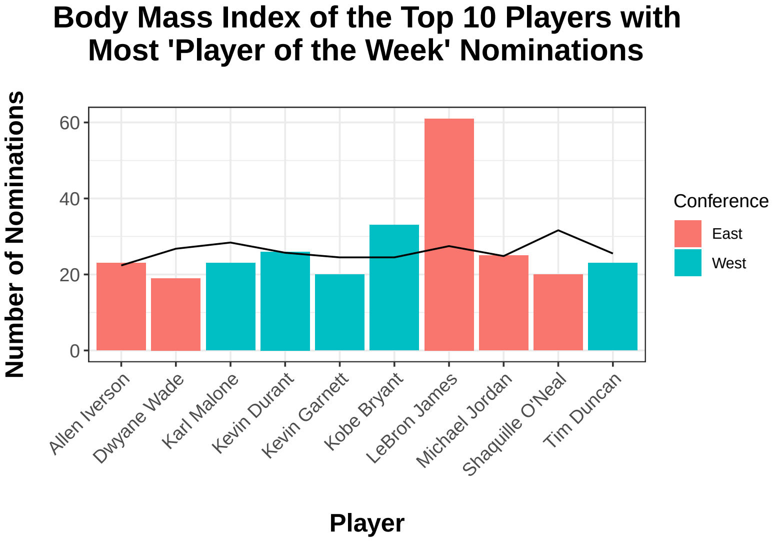
```
[1] 1.9558
```

```
# A tibble: 10 x 6
   Player           Height Weight Conference     n   BMI
   <fct>             <dbl>  <dbl> <fct>      <int> <dbl>
 1 LeBron James       2.03  113.  East          61  27.5
 2 Kobe Bryant        1.98   96.2 West          33  24.5
 3 Kevin Durant       2.06  109.  West          26  25.7
 4 Michael Jordan     1.98   97.5 East          25  24.8
 5 Allen Iverson      1.83   74.8 East          23  22.4
 6 Karl Malone        2.06  120.  West          23  28.4
 7 Tim Duncan         2.11  113.  West          23  25.5
 8 Kevin Garnett      2.11  109.  West          20  24.5
 9 Shaquille O'Neal   2.16  147.  East          20  31.6
10 Dwyane Wade        1.93   99.8 East          19  26.8
```

Then, I used the `pairs()` function to analyze the correlation between BMI and the number of nominations to the Player of The Week award.



After formatting the data according to get the BMI of each of the players, I plotted the BMI of of the top 10 players with most Player of The Week nominations.



The results showed that physical attributes such as weight and height (and consequently BMI) do not have a direct impact on the number of nominations to the Player of The Week award among the players with more than one nomination. However, it is possible to analyze that the winners have on average 99 kilos and 2 meters of height. In this sense, on average they have a BMI considered normal (24.5).