440 Project

Team Name - Flyin' Illini Data Dorks - Clare Oehler-O'Sullivan, Vishesh Gupta, Aneesh Balusu

Title - Basketball Data Management

Introduction

Is there a recipe for success in the NBA? Will going to a certain college, having a certain weight to height ratio, etc. make you more likely to succeed? We will be exploring the success of players with various different backgrounds in the NBA and seeing if there are any correlations.

Questions to Address:

1. Which universities tend to create athletes who are drafted into the NBA the most?
2. Who are the top scoring players and where did they come from?
3. Is there a correlation between different positions and what height the players of those positions should be?
4. Do the top scoring players have a higher weight to height ratio or lower than the mean weight-height ratio?
5. Which positions are most likely to be NBA Player of the Year?

Research Interest

We are interested in creating a combined table using either 2 or 3 of the below data files to seek further information regarding the correlation between the transition of players from a university level to the NBA level. We intended on using the lower 2 mentioned data sets to try and summarize which universities tend to create athletes who are drafted into the NBA and have been selected in the all star team on a yearly basis. We are then interested in making predictions from the first basketball data set to understand which positions are most likely to be NBA Player of the Year and seeing if there is a correlation between # of Seasons in League and Age. We could further summarise the data to look at which variables in the combined data set are redundant to our response variable and then summarise the remaining predictors.

We will first merge our data sets to get a collective data sets with the required observations. Based on our initial observation this will be done using the players name. This combined data set will consist of multiple inconsistencies as we will have to remove rows consisting of NA's in variables such as college, team, pos and so on. We will possibly have to ensure that the players teams are well documented as one player could have played for multiple teams over his career. Therefore we will need to convert the Team abbreviations to their full name in our combined data set to ensure we have merged the data correctly. We may need to make various data type changes such as converting the height column into inches, the date column into date type and more.

Methods

Data files:

1. NBA Player of the Week dataset : https://urldefense.com/v3/___https://www.kaggle.com/jacobbaruch/nba-player-of-the-week___;!!DZ3fjg!tnr4cfFS_MOGsYi4NFFmQ-xVGf-Y8Dzbc5pRMjpjv_V8I83zavnaf7

- variables: 17
- observations: 1,345

Some important columns include: Player, Team, Position, Age, Draft Year, # of Seasons in League.

This dataset got its data from this source: https://urldefense.com/v3/___https://basketball.realgm.com/___;!!DZ3fjg!tnr4cfFS_MOGsYi4NFFmQ-xVGf-Y8Dzbc5pRMjpjv_V8I83zavnaf7GL3vCKh1xwKhNG$ Citation: "Basketball News, Rumors, Scores, Stats, Analysis, Depth Charts, Forums." RealGM, basketball.realgm.com/.

2. Two decades of data on each player who has been part of an NBA teams' roster: https://urldefense.com/v3/___https://www.kaggle.com/justinas/nba-players-data___;!!DZ3fjg!tnr4cfFS_MOGsYi4NFFmQ-xVGf-Y8Dzbc5pRMjpjv_V8I83zavnaf7GL3vCKh1Bb9NbU$

- variable: 22
- observation: 11146

Some important columns across Two decades of data on each player who has been part of an NBA teams' roster include: player_name, team_abbreviation, college, age, draft year, pts

3. NBA all star teams: https://urldefense.com/v3/___https://www.kaggle.com/fmejia21/nba-all-star-game-20002016___;!!DZ3fjg!tnr4cfFS_MOGsYi4NFFmQ-xVGf-Y8Dzbc5pRMjpjv_V8I83zavnaf7GL3vCKh_4G-IX-$

- variable: 9
- observation: 439

Background information on these Files:

The NBA Player of the Week dataset got its data from this source:

https://urldefense.com/v3/___https://basketball.realgm.com/___;!!DZ3fjg!tUll3tseX9zHytgVMYzqdCwGsk1-O5CaIzSwoS8CE7BGNgu6ml111t1I20XCHkDstL0z$

Citation: "Basketball News, Rumors, Scores, Stats, Analysis, Depth Charts, Forums." RealGM, basketball.realgm.com/.

The NBA Player all star teams dataset got its data from this source:

https://urldefense.com/v3/___https://www.kaggle.com/fmejia21/nba-all-star-game-20002016___;!!DZ3fjg!tnr4cfFS_MOGs
xVGf-Y8Dzbc5pRMjpjv_V8I83zavnaf7GL3vCKh_4G-IX-$

The data set regarding two decades of data on each player who has been part of an NBA teams' roster got its data from this source:

https://urldefense.com/v3/___https://www.kaggle.com/justinas/nba-players-data___;!!DZ3fjg!tnr4cfFS_MOGsYi4NFFmQ
xVGf-Y8Dzbc5pRMjpjv_V8I83zavnaf7GL3vCKh1Bb9NbU$

Results

Loading Libraries

We will be using the library tidyverse, because it contains many helpful functions that we commonly use in R code. We will be using the library readr to read our data in from their sources from csv format into a dataframe. We will be using the library sqldf to write SQL code in R. We will be using dpylr because this library makes it easier to do multiple operations on a single dataset easier.

```r
library(tidyverse)
library(readr)
library(sqldf)
library(dplyr)
```

Loading Data and formatting date columns to be date format

On uploading the data we noticed that for two of our data sets the dates data type was set as a character. Therefore, we decided to convert the data type to a date and formatted the information for each data set in the patter of month-day-year

```r
NBA_player_of_the_week <- read_csv("NBA_player_of_the_week.csv",
    col_types = cols(Date = col_date(format = "%b %d, %Y")))


all_seasons <- read_csv("all_seasons.csv")
```

Merge these NBA datasets and making it so there is only one occurance of each player and the data on both of these seasons about the player

While we are combining both the data sets we need to ensure that the names of the players are unique i.e. there are no multiple occurrences of a players name. Therefore when we are creating 'remove_dup_players_all_seasons' we group by player_name. To then have a common unique variable which can be used to combine the two data sets we need to replace 'player_name' as 'Player' so that an innew join can be performed.

```r
length(unique(all_seasons$player_name))
```

```
## [1] 2235
```

```r
# ensure our data only contains unique player names and no duplicates for each data file read in

remove_dup_players_all_seasons =sqldf("  SELECT *
                                      FROM   all_seasons
                                      GROUP BY player_name
                                      ")

remove_dup_players_all_seasons = remove_dup_players_all_seasons %>%
  mutate (Player = player_name)

remove_dup_players_NBA_player_of_the_week = sqldf("
  SELECT *
  FROM   NBA_player_of_the_week
  GROUP BY Player
  ")

#create one common data set
combined_unique_player_set = inner_join(remove_dup_players_all_seasons,
                                        remove_dup_players_NBA_player_of_the_week,
                                        by = 'Player')

str(combined_unique_player_set)
```

```
## 'data.frame':    251 obs. of  39 variables:
##  $ X1               : num  139 1037 5032 3820 146 ...
##  $ player_name      : chr  "Aaron McKie" "Al Harrington" "Al Horford" "Al Jefferson" ...
##  $ team_abbreviation: chr  "DET" "IND" "ATL" "BOS" ...
##  $ age              : num  24 19 22 20 26 22 27 20 19 24 ...
##  $ player_height    : num  196 206 208 208 198 ...
##  $ player_weight    : num  94.8 104.3 111.1 120.2 90.7 ...
##  $ college          : chr  "Temple" "None" "Florida" "None" ...
##  $ country          : chr  "USA" "USA" "Dominican Republic" "USA" ...
##  $ draft_year       : chr  "1994" "1998" "2007" "2004" ...
##  $ draft_round      : chr  "1" "1" "1" "1" ...
##  $ draft_number     : chr  "17" "25" "3" "15" ...
##  $ gp               : num  83 21 81 71 81 76 66 82 60 82 ...
##  $ pts              : num  5.2 2.1 10.1 6.7 14.8 23.5 19.8 13.5 7.9 11.1 ...
##  $ reb              : num  2.7 1.9 9.7 4.4 3 4.1 9.9 8.8 7.6 3.4 ...
##  $ ast              : num  1.9 0.2 1.5 0.3 2.2 7.5 1.6 1 0.5 5.8 ...
##  $ net_rating       : num  5.2 -8.3 -2.4 -1.7 2.9 -7 10.5 -0.8 -1.5 -1.6 ...
##  $ oreb_pct         : num  0.031 0.132 0.113 0.137 0.02 0.04 0.1 0.109 0.153 0.046 ...
##  $ dreb_pct         : num  0.129 0.148 0.247 0.205 0.086 0.072 0.229 0.206 0.266 0.105 ...
##  $ usg_pct          : num  0.147 0.21 0.162 0.212 0.223 0.284 0.275 0.214 0.17 0.222 ...
##  $ ts_pct           : num  0.524 0.359 0.539 0.554 0.531 0.513 0.578 0.53 0.578 0.517 ...
##  $ ast_pct          : num  0.163 0.06 0.078 0.042 0.117 0.32 0.086 0.05 0.04 0.386 ...
##  $ season           : chr  "1996-97" "1998-99" "2007-08" "2004-05" ...
##  $ Player           : chr  "Aaron McKie" "Al Harrington" "Al Horford" "Al Jefferson" ...
##  $ Team             : chr  "Philadelphia Sixers" "New York Knicks" "Atlanta Hawks" "Charlotte Bobcats
##  $ Conference       : chr  NA "East" "East" "East" ...
##  $ Date             : Date, format: "2000-12-31" "2008-12-15" ...
##  $ Position         : chr  "G" "F" "FC" "FC" ...
##  $ Height           : chr  "6'5" "6'9" "6'10" "6'10" ...
##  $ Weight           : num  209 245 245 289 205 165 240 245 279 200 ...
##  $ Age              : num  28 29 28 29 31 32 30 28 25 31 ...
##  $ Draft Year       : num  1994 1998 2007 2004 1993 ...
##  $ Seasons in league: num  6 10 7 9 9 11 7 8 6 8 ...
##  $ Season           : chr  "2000-2001" "2008-2009" "2014-2015" "2013-2014" ...
##  $ Season short     : num  2001 2009 2015 2014 2003 ...
##  $ Pre-draft Team   : chr  "Temple" "St. Patrick High School (New Jersey)" "Florida" "Prentiss High S
##  $ Real_value       : num  1 0.5 0.5 0.5 0.5 0.5 1 0.5 0.5 0.5 ...
##  $ Height CM        : num  196 206 208 208 198 183 208 208 211 190 ...
##  $ Weight KG        : num  94 111 111 131 93 74 108 111 126 90 ...
##  $ Last Season      : num  0 0 0 0 0 0 0 0 0 0 ...
```

Getting rid of duplicates of the same column name

When we merged these datasets, we can observe that some of the columns in these datasets have the same name with a different capitalization, such as 'Age' and 'age'. R does not handle having multiple of the same column name well so we went in and renamed the age and season from the Player of the Week dataset to be new_age and new_season that way there is no confusion between multiple of the same column name.

We were able to do this by creating a new column name called new_age and new_season and assigning these column names to contain the data in Age and Season, respectively. From here, we dropped the old columns Age and Season.

```
combined_unique_player_set = combined_unique_player_set %>%
  mutate(new_age = Age) %>%
  select(-Age) %>%
  mutate(new_season = Season) %>%
  select(-Season)
```

Ordering these players in descending order by the number of points they score per game

We want to order these players by the average number of points they score per game that way we can determine who the best shooters are. Then we will be able to see what schools the highest scoring players are from. We are interested in comparing these colleges with the colleges who have the players with the most assists because we are caring about the best offensive teams. We would like to see if some of the same schools have the players with the most points scored and the most assists.

We will output the 10 highest scoring players in this dataset because right now we are only looking at the very best players.

```
scorers = sqldf("
                                    SELECT Player, team_abbreviation, college, draft_year, net_rating, p
                                    FROM combined_unique_player_set
                                    ORDER BY pts desc
                                    ")

top_scorers = head(scorers, 10)
top_scorers
```

```
##              Player team_abbreviation         college draft_year net_rating
## 1    Michael Jordan               CHI  North Carolina       1984       13.4
## 2       Karl Malone               UTA  Louisiana Tech       1985       12.8
## 3         Glen Rice               CHH        Michigan       1989        3.2
## 4  Shaquille O'Neal               LAL Louisiana State       1992        6.9
## 5    Mitch Richmond               SAC    Kansas State       1988       -2.0
## 6     Allen Iverson               PHI      Georgetown       1996       -7.0
## 7   Hakeem Olajuwon               HOU         Houston       1984        6.5
## 8     Blake Griffin               LAC        Oklahoma       2009       -3.4
## 9     Patrick Ewing               NYK      Georgetown       1985        6.4
## 10      Gary Payton               SEA    Oregon State       1990       10.3
##     pts Position
## 1  29.6       SG
## 2  27.4       PF
## 3  26.8       SF
## 4  26.2        C
## 5  25.9       SG
## 6  23.5        G
## 7  23.2        C
## 8  22.5       PF
## 9  22.4        C
## 10 21.8        G
```

Ordering these players in descending order by their assist percentage over all seasons

For anyone who is unfamiliar with basketball terms, assist percentage refers to the percentage of field goals a teammate made that this player passed the ball to. A player's assist percentage can tell you a lot about how well a player knows the game and is able to play offensively.

We will do the same thing that we did for average points scored per game to the players assist percentage, ordering in descending order by assist percentage. We will output the 10 highest assisting players in this dataset and compare these to the 10 highest scoring players. We will be able to see if these are any players in common between the top scorers and the top assisters. We are only selecting to output the relevant columns.

```
assisters = sqldf("
                                  SELECT Player, team_abbreviation, college, draft_year, net_rating, ast
                                   FROM combined_unique_player_set
                                   ORDER BY ast_pct desc
                                   ")

top_assisters = head(assisters, 10)
top_assisters
```

```
##                Player team_abbreviation         college draft_year net_rating
## 1        Mark Jackson               IND St. John's (NY)       1987       -2.0
## 2       John Stockton               UTA         Gonzaga       1984       11.4
## 3        Tim Hardaway               MIA   Texas-El Paso       1989        8.9
## 4   Spencer Dinwiddie               DET        Colorado       2014       -8.7
## 5          Trae Young               ATL        Oklahoma       2018       -6.3
## 6       Kevin Johnson               PHX      California       1987        2.9
## 7        Andre Miller               CLE            Utah       1999       -1.6
## 8          Chris Paul               NOK     Wake Forest       2005       -3.9
## 9            T.J. Ford               MIL           Texas       2003       -0.2
## 10     Stephon Marbury              MIN    Georgia Tech       1996       -3.1
##    ast_pct Position
## 1    0.464       PG
## 2    0.450       PG
## 3    0.404       PG
## 4    0.394       PG
## 5    0.390       PG
## 6    0.388       PG
## 7    0.386       PG
## 8    0.378       PG
## 9    0.375       PG
## 10   0.373        G
```

We can observe that there are no matches in the data of players who are both top scorers and top assisters. We can conclude that different players are good at scoring and assisting.

But now, we are curious about which college programs are the best at offense collectively and have had the best scorers and assisters.

Making a frequency table for the Colleges that have players in the top 50 Scorers & top 50 Assisters.

In order to see which college programs have had the best scorers, we will select the first 50 entries of the sorted data on the the average scoring percentage.

To do this correctly, we need to filter out colleges listed as 'None'. There will be 'None' in the data under college for any players that did not go to college and joined the NBA right out of high school. The way we will determine the frequency of each college program in the top 50 scoring players, we will count each program by the number of times it appears in the top 50 dataset. Then, we will sort the data in descending order of their frequencies and alphabetically by college name after that.

```
top_50_scorers = scorers[1:50,]

top_scoring_programs = sqldf("
      select college, count(college) as frequency
      from top_50_scorers
      where college not in ('None')
      group by college
      order by frequency desc, college
      ")

best_scoring = head(top_scoring_programs, 10)
best_scoring
```

```
##           college frequency
## 1            Duke         3
## 2      Georgetown         3
## 3        Michigan         3
## 4  North Carolina         3
## 5        Oklahoma         3
## 6      California         2
## 7         Houston         2
## 8        Kentucky         2
## 9        Syracuse         2
## 10        Alabama         1
```

Now, we will take a subset of the top 50 assisting players from the sorted assisters dataset. We are able to simply take the first 50 entries of this dataset because it is already sorted by assist percentage in descending order.

Again, we filtered out the colleges in this dataset that were labelled 'None'. We counted the number of times a college was in the dataset using the count function and setting that to a new column name, frequency. We grouped our data by college with the group by function so that each program is only listed once. Finally, we sorted in descending order by frequency and then alphabetically by college.

```
top_50_assisters = assisters[1:50,]

top_assisting_programs = sqldf("
      select college, count(college) as frequency
      from top_50_assisters
      where college not in ('None')
      group by college
      order by frequency desc, college
      ")

best_assisting = head(top_assisting_programs, 10)
best_assisting
```

```
##         college frequency
## 1       Arizona         2
## 2    California         2
```

```
## 3          Duke         2
## 4  Georgia Tech         2
## 5      Kentucky         2
## 6      Maryland         2
## 7       Memphis         2
## 8      Oklahoma         2
## 9          UCLA         2
## 10  Wake Forest         2
```

Which programs are in the top 10 list for both top scoring programs and top assisting programs?

I iterated through each entry in the top 10 best scoring colleges and for each of those college I checked if the college was the same as each of the top 10 best scoring colleges. Each time I found a match, I printed out the college.

```
for (i in 1:nrow(best_scoring)){
  for (j in 1:nrow(best_assisting)){
    if (best_scoring$college[i] == best_assisting$college[j]){
      print(best_scoring$college[i])
    }
  }
}
```

```
## [1] "Duke"
## [1] "Oklahoma"
## [1] "California"
## [1] "Kentucky"
```

Generally speaking, the players at Duke, Oklahoma, California, and Kentucky have had the most top notch scoring and assisting players out of this dataset and these schools have some the highest quality offensive programs. If a top-notch player who wants to enter the NBA wants to have the best chances of success at scoring or assisting, he should consider going to college at Duke, Oklahoma, California, or Kentucky because these schools train their players well offensively and these players go on to be top scoreres and assisters in the NBA.

Now, we will explore a new topic related to the success of NBA Players, their weight to height ratio.

Many players often wonder whether it is more beneficial to be bulky with muscle for basketball or to be lean. The right balance between bulking up and slimming down is different for everyone because we all have different body types. However, we are curious to explore whether players who have a higher weight to height ratio than the average NBA player are more successful or if a lower weight to height ratio is more successful on average.

To do this, we will add a new variable called wh_ratio that is the Weight to Height ratio of the players. We will compare the ratio for the average player on the team to each of the top scorers and see if the top scorers have more or less weight than the average ratio.

We will output our dataframe that contains applicable columns and the average weight to height ratio. In this data, the variable Weight is in kg and height is in cm.

```r
combined_unique_player_set = combined_unique_player_set %>%
  mutate(wh_ratio = player_weight / player_height)

weight_height_ratio_set = sqldf("
                                  SELECT Player, team_abbreviation, college, draft_year, net_rating, a
                                  FROM combined_unique_player_set
                                  ORDER BY wh_ratio desc
                                  ")
head(weight_height_ratio_set, 10)
```

```
##                Player team_abbreviation          college draft_year net_rating
## 1      Oliver Miller               TOR         Arkansas       1992       -1.0
## 2  Shaquille O'Neal               LAL  Louisiana State       1992        6.9
## 3      Carlos Boozer               CLE             Duke       2002      -11.9
## 4    Arvydas Sabonis               POR             None       1986        8.1
## 5      Zach Randolph               POR   Michigan State       2001       -3.7
## 6      Larry Johnson               NYK Nevada-Las Vegas       1991        4.7
## 7           Yao Ming               HOU             None       2002        2.2
## 8     Andre Drummond               DET      Connecticut       2012       -1.5
## 9       Andrew Bynum               LAL             None       2005       -4.0
## 10 DeMarcus Cousins               SAC         Kentucky       2010       -7.5
##    ast_pct  wh_ratio
## 1    0.116 0.6834525
## 2    0.159 0.6302807
## 3    0.090 0.6173119
## 4    0.136 0.5993704
## 5    0.103 0.5952651
## 6    0.115 0.5945116
## 7    0.104 0.5939274
## 8    0.040 0.5880058
## 9    0.044 0.5846354
## 10   0.149 0.5809214
```

```r
avg_ratio = mean(weight_height_ratio_set$wh_ratio)
avg_ratio
```

```
## [1] 0.4943594
```

Do the top scoring players have a higher weight to height ratio or lower than the mean weight-height ratio?

We will now add another column to this dataset called higher_ratio that returns TRUE when a player's weight to height ratio is higher than the average weight to height ratio and FALSE when it is lower.

```r
score_ratio_df = sqldf("
      select top_scorers.Player, top_scorers.pts, weight_height_ratio_set.wh_ratio
      from weight_height_ratio_set, top_scorers
      where top_scorers.Player = weight_height_ratio_set.Player
      order by pts desc
      ")
```

```
score_ratio_df$higher_ratio = ifelse(score_ratio_df$wh_ratio > avg_ratio, TRUE, FALSE)
score_ratio_df
```

```
##               Player  pts  wh_ratio higher_ratio
## 1    Michael Jordan 29.6 0.4945279         TRUE
## 2       Karl Malone 27.4 0.5643995         TRUE
## 3         Glen Rice 26.8 0.4910937        FALSE
## 4  Shaquille O'Neal 26.2 0.6302807         TRUE
## 5    Mitch Richmond 25.9 0.4986311         TRUE
## 6     Allen Iverson 23.5 0.4092448        FALSE
## 7   Hakeem Olajuwon 23.2 0.5421164         TRUE
## 8     Blake Griffin 22.5 0.5466276         TRUE
## 9     Patrick Ewing 22.4 0.5102272         TRUE
## 10      Gary Payton 21.8 0.4464488        FALSE
```

7/10 of the top scorers are above average in their weight to height ratio. According to our data, on average, the highest scoring basketball players are the ones who weigh more. This could be due to their muscle mass.

Do the top assisting players have a higher weight to height ratio or lower than the mean weight-height ratio?

We will also add another column to this dataset called higher_ratio that returns TRUE when a player's weight to height ratio is higher than the average weight to height ratio and FALSE when it is lower.

```
ast_ratio_df = sqldf("
      select top_assisters.Player, top_assisters.ast_pct, weight_height_ratio_set.wh_ratio
      from weight_height_ratio_set, top_assisters
      where top_assisters.Player = weight_height_ratio_set.Player
      order by top_assisters.ast_pct desc
      ")

ast_ratio_df$higher_ratio = ifelse(ast_ratio_df$wh_ratio > avg_ratio, TRUE, FALSE)
ast_ratio_df
```

```
##                Player ast_pct  wh_ratio higher_ratio
## 1        Mark Jackson   0.464 0.4404962        FALSE
## 2       John Stockton   0.450 0.4281016        FALSE
## 3        Tim Hardaway   0.404 0.4836529        FALSE
## 4  Spencer Dinwiddie   0.394 0.4578962        FALSE
## 5          Trae Young   0.390 0.4343826        FALSE
## 6       Kevin Johnson   0.388 0.4647960        FALSE
## 7        Andre Miller   0.386 0.4898871        FALSE
## 8          Chris Paul   0.378 0.4340475        FALSE
## 9           T.J. Ford   0.375 0.4092448        FALSE
## 10    Stephon Marbury   0.373 0.4343826        FALSE
```

7/10 of the top assisters also are above average in their weight to height ratio. According to our data, on average, the highest assisting basketball players are the ones who weigh more. This might also be due to their muscle mass.

We now used a common methodology amongst all the three cleaned datasets above along with the new formed joined data. To obtain the best univeristy program we:

- Removed all the NA values
- Created a frequency table of the dataset being taken into consideration
- Calculated the percentage of each unique college
- Obtained the best program by finding the highest percentage

Based on the answers we ensure that the players within that subcategory are distinct and then based on this conclusion we further looked at what are the necessary parameters for success while you are still in the program.

Based on the all seasons data only:

```
# create a frequency table to understand the university with the highest percentage of student athletes

number_of_players_1 = table(remove_dup_players_all_seasons$college)
number_of_players_1 = as.data.frame(number_of_players_1)
number_of_players_1 = number_of_players_1%>%
  mutate(per_per_school = (number_of_players_1$Freq/nrow(number_of_players_1))*100)
x1 = sort(number_of_players_1$per_per_school)
x1 = x1[306]
z1 = (number_of_players_1$per_per_school == x1)
number_of_players_1$Var1[z1]
```

```
## [1] Kentucky
## 307 Levels: Alabama Alabama A&M Alabama Huntsville ... Yonsei (KOR)
```

```
#statistical analysis
#college decided based on above code chucks observation
players_from_best_uni_1=sqldf("
  SELECT *
  FROM   remove_dup_players_all_seasons
  WHERE  college == 'Kentucky'
  ")
players_from_best_uni_1
```

```
##       X1           player_name team_abbreviation age player_height
## 1   8924        Aaron Harrison               CHA  21        198.12
## 2   9090        Alex Poythress               PHI  23        200.66
## 3   9080       Andrew Harrison               MEM  22        198.12
## 4   7556         Anthony Davis               NOH  20        208.28
## 5    159        Antoine Walker               BOS  20        205.74
## 6   8067        Archie Goodwin               PHX  19        195.58
## 7   9616           Bam Adebayo               MIA  20        208.28
## 8   7089        Brandon Knight               DET  20        190.50
## 9   4248           Chuck Hayes               HOU  23        198.12
## 10  9745        Dakari Johnson               OKC  22        213.36
## 11  6780          Daniel Orton               ORL  21        208.28
## 12  7479         Darius Miller               NOH  23        203.20
## 13  9686          De'Aaron Fox               SAC  20        190.50
## 14  6782       DeAndre Liggins               ORL  24        198.12
## 15  6456      DeMarcus Cousins               SAC  20        210.82
## 16   711         Derek Anderson               CLE  23        195.58
```

```
## 17  8870            Devin Booker               PHX  19        198.12
## 18  7547              Doron Lamb               ORL  21        193.04
## 19  6681             Enes Kanter               UTA  20        210.82
## 20  6334            Eric Bledsoe               LAC  21        185.42
## 21  3965            Erik Daniels               SAC  23        203.20
## 22  4374            Gerald Fitch               MIA  23        190.50
## 23 10116          Hamidou Diallo               OKC  20        195.58
## 24 10127         Isaac Humphries               ATL  21        213.36
## 25  1817         Jamaal Magloire               CHH  23        208.28
## 26    65          Jamal Mashburn               MIA  24        203.20
## 27  9252            Jamal Murray               DEN  20        193.04
## 28  8393             James Young               BOS  19        198.12
## 29 10197       Jarred Vanderbilt               DEN  20        205.74
## 30  1315           Jeff Sheppard               ATL  24        190.50
## 31  5838             Jodie Meeks               PHI  22        193.04
## 32  5543            Joe Crawford               NYK  23        195.58
## 33  6285               John Wall               WAS  20        193.04
## 34  7005         Josh Harrellson               NYK  23        208.28
## 35  8281           Julius Randle               LAL  20        205.74
## 36  8781      Karl-Anthony Towns               MIN  20        213.36
## 37  3285            Keith Bogans               ORL  24        195.58
## 38 11073          Keldon Johnson               SAS  20        195.58
## 39  4812        Kelenna Azubuike               GSW  23        195.58
## 40 10496           Kevin Knox II               NYK  19        205.74
## 41 10036              Malik Monk               CHA  20        190.50
## 42   464               Mark Pope               IND  25        208.28
## 43  7203          Marquis Teague               CHI  20        187.96
## 44  7232  Michael Kidd-Gilchrist               CHA  19        200.66
## 45 11096           Mychal Mulder               GSW  25        190.50
## 46  1124           Nazr Mohammed               PHI  21        208.28
## 47  8582            Nerlens Noel               PHI  21        210.82
## 48 10924         P.J. Washington               CHA  21        200.66
## 49  6621       Patrick Patterson               HOU  22        205.74
## 50  4696             Rajon Rondo               BOS  21        185.42
## 51  4697         Randolph Morris               NYK  21        210.82
## 52   629           Reggie Hanson               BOS  29        203.20
## 53   262             Rex Chapman               PHX  29        193.04
## 54   612              Ron Mercer               BOS  22        200.66
## 55  1548           Scott Padgett               UTA  24        205.74
## 56 10372 Shai Gilgeous-Alexander               LAC  20        198.12
## 57  9321          Skal Labissiere               SAC  21        210.82
## 58  2938          Tayshaun Prince               DET  23        205.74
## 59  7296          Terrence Jones               HOU  21        205.74
## 60   305               Tony Delk               CHH  23        187.96
## 61  8980              Trey Lyles               UTA  20        208.28
## 62 10966             Tyler Herro               MIA  20        195.58
## 63  9139              Tyler Ulis               PHX  21        177.80
## 64   293          Walter McCarty               NYK  23        208.28
## 65  1614             Wayne Turner               BOS  24        187.96
## 66  9058      Willie Cauley-Stein               SAC  22        213.36
##    player_weight  college   country draft_year draft_round draft_number gp  pts
## 1       95.25432 Kentucky       USA  Undrafted   Undrafted    Undrafted 21  0.9
## 2      107.95490 Kentucky       USA  Undrafted   Undrafted    Undrafted  6 10.7
## 3       96.61510 Kentucky       USA       2015           2           44 72  5.9
```

```
## 4       99.79024 Kentucky       USA       2012           1            1 64 13.5
## 5      101.60461 Kentucky       USA       1996           1            6 82 17.5
## 6       89.81122 Kentucky       USA       2013           1           29 52  3.7
## 7      115.66596 Kentucky       USA       2017           1           14 69  6.9
## 8       85.72889 Kentucky       USA       2011           1            8 66 12.8
## 9      109.76926 Kentucky       USA  Undrafted   Undrafted    Undrafted 40  3.7
## 10     115.66596 Kentucky       USA       2015           2           48 31  1.8
## 11     115.66596 Kentucky       USA       2010           1           29 16  2.8
## 12     106.59412 Kentucky       USA       2012           2           46 52  2.3
## 13      79.37860 Kentucky       USA       2017           1            5 73 11.6
## 14      94.80073 Kentucky       USA       2011           2           53 17  1.9
## 15     122.46984 Kentucky       USA       2010           1            5 81 14.1
## 16      88.45044 Kentucky       USA       1997           1           13 66 11.7
## 17      93.43995 Kentucky       USA       2015           1           13 76 13.8
## 18      95.25432 Kentucky       USA       2012           2           42 47  3.3
## 19     121.10906 Kentucky    Turkey       2011           1            3 66  4.6
## 20      88.45044 Kentucky       USA       2010           1           18 81  6.7
## 21      97.06869 Kentucky       USA  Undrafted   Undrafted    Undrafted 21  0.6
## 22      85.27530 Kentucky       USA  Undrafted   Undrafted    Undrafted 18  4.7
## 23      89.81122 Kentucky       USA       2018           2           45 51  3.7
## 24     117.93392 Kentucky Australia  Undrafted   Undrafted    Undrafted  5  3.0
## 25     117.93392 Kentucky    Canada       2000           1           19 74  4.6
## 26     113.39800 Kentucky       USA       1993           1            4 69 11.9
## 27      93.89354 Kentucky    Canada       2016           1            7 82  9.9
## 28      97.52228 Kentucky       USA       2014           1           17 31  3.4
## 29      97.06869 Kentucky       USA       2018           2           41 17  1.4
## 30      86.18248 Kentucky       USA  Undrafted   Undrafted    Undrafted 17  2.4
## 31      94.34714 Kentucky       USA       2009           2           41 60  4.7
## 32      95.25432 Kentucky       USA       2008           2           58  2  4.5
## 33      88.45044 Kentucky       USA       2010           1            1 69 16.4
## 34     124.73780 Kentucky       USA       2011           2           45 37  4.4
## 35     113.39800 Kentucky       USA       2014           1            7  1  2.0
## 36     110.67645 Kentucky       USA       2015           1            1 82 18.3
## 37      97.52228 Kentucky       USA       2003           2           43 73  6.8
## 38      99.79024 Kentucky       USA       2019           1           29  7  4.7
## 39      99.79024 Kentucky   England  Undrafted   Undrafted    Undrafted 41  7.1
## 40      97.52228 Kentucky       USA       2018           1            9 75 12.8
## 41      90.71840 Kentucky       USA       2017           1           11 63  6.7
## 42     106.59412 Kentucky       USA       1996           2           52 28  1.4
## 43      86.18248 Kentucky       USA       2012           1           29 48  2.1
## 44     105.23334 Kentucky       USA       2012           1            2 78  9.0
## 45      83.46093 Kentucky    Canada  Undrafted   Undrafted    Undrafted  6 12.3
## 46     108.86208 Kentucky       USA       1998           1           29 26  1.6
## 47     103.41898 Kentucky       USA       2013           1            6 75  9.9
## 48     104.32616 Kentucky       USA       2019           1           12 56 12.3
## 49     106.59412 Kentucky       USA       2010           1           14 52  6.3
## 50      77.56423 Kentucky       USA       2006           1           21 78  6.4
## 51     117.93392 Kentucky       USA  Undrafted   Undrafted    Undrafted  5  0.8
## 52      88.45044 Kentucky       USA       1998   Undrafted    Undrafted  8  0.8
## 53      88.45044 Kentucky       USA       1988           1            8 65 13.8
## 54      95.25432 Kentucky       USA       1997           1            6 80 15.3
## 55     108.86208 Kentucky       USA       1999           1           28 47  2.6
## 56      82.10015 Kentucky    Canada       2018           1           11 82 10.8
## 57     102.05820 Kentucky     Haiti       2016           1           28 33  8.8
```

```
## 58      97.52228 Kentucky       USA       2002           1            23 42  3.3
## 59     114.30518 Kentucky       USA       2012           1            18 19  5.5
## 60      85.72889 Kentucky       USA       1996           1            16 61  5.4
## 61     106.14053 Kentucky       USA       2015           1            12 80  6.1
## 62      88.45044 Kentucky       USA       2019           1            13 46 13.1
## 63      68.03880 Kentucky       USA       2016           2            34 61  7.3
## 64     104.32616 Kentucky       USA       1996           1            19 35  1.8
## 65      86.18248 Kentucky       USA  Undrafted   Undrafted    Undrafted  3  1.3
## 66     108.86208 Kentucky       USA       2015           1             6 66  7.0
##     reb ast net_rating oreb_pct dreb_pct usg_pct ts_pct ast_pct  season
## 1   0.7 0.1        2.2    0.047    0.133   0.138  0.371   0.033 2015-16
## 2   4.8 0.8       -9.6    0.073    0.137   0.166  0.548   0.052 2016-17
## 3   1.9 2.8       -0.3    0.017    0.090   0.164  0.477   0.209 2016-17
## 4   8.2 1.0       -5.5    0.106    0.236   0.215  0.559   0.060 2012-13
## 5   9.0 3.2       -9.3    0.103    0.189   0.251  0.474   0.150 1996-97
## 6   1.7 0.4       -7.6    0.051    0.121   0.192  0.507   0.061 2013-14
## 7   5.5 1.5       -0.6    0.086    0.194   0.157  0.570   0.115 2017-18
## 8   3.2 3.8       -7.2    0.018    0.103   0.217  0.511   0.202 2011-12
## 9   4.5 0.4        9.2    0.140    0.253   0.120  0.589   0.044 2005-06
## 10  1.1 0.3       10.9    0.099    0.111   0.139  0.575   0.083 2017-18
## 11  2.4 0.3      -11.5    0.109    0.130   0.120  0.549   0.047 2011-12
## 12  1.5 0.8       -5.1    0.013    0.123   0.092  0.529   0.100 2012-13
## 13  2.8 4.4      -10.2    0.016    0.087   0.227  0.478   0.240 2017-18
## 14  0.9 0.3      -27.1    0.058    0.103   0.182  0.495   0.091 2011-12
## 15  8.6 2.5       -7.5    0.105    0.247   0.272  0.484   0.149 2010-11
## 16  2.8 3.4        3.9    0.038    0.091   0.215  0.531   0.219 1997-98
## 17  2.5 2.6       -9.6    0.014    0.083   0.231  0.535   0.162 2015-16
## 18  1.0 0.7       -8.9    0.016    0.075   0.156  0.433   0.085 2012-13
## 19  4.2 0.1       -8.2    0.135    0.231   0.169  0.539   0.016 2011-12
## 20  2.8 3.6       -7.0    0.045    0.098   0.179  0.499   0.254 2010-11
## 21  0.9 0.2      -23.2    0.090    0.162   0.141  0.361   0.118 2004-05
## 22  1.7 1.8        0.3    0.033    0.105   0.216  0.424   0.228 2005-06
## 23  1.9 0.3       -6.8    0.065    0.110   0.160  0.497   0.044 2018-19
## 24  2.2 0.0      -17.1    0.055    0.123   0.142  0.357   0.000 2018-19
## 25  4.0 0.4        3.2    0.111    0.195   0.164  0.506   0.042 2000-01
## 26  4.3 2.9        1.4    0.037    0.121   0.207  0.487   0.158 1996-97
## 27  2.6 2.1        1.0    0.026    0.104   0.216  0.518   0.142 2016-17
## 28  1.4 0.4       -2.9    0.027    0.110   0.153  0.457   0.059 2014-15
## 29  1.4 0.2        4.5    0.084    0.205   0.179  0.513   0.064 2018-19
## 30  1.3 0.9       -2.9    0.036    0.098   0.130  0.447   0.157 1998-99
## 31  1.7 0.7       -3.1    0.014    0.154   0.190  0.493   0.080 2009-10
## 32  2.0 0.5       16.8    0.043    0.120   0.216  0.414   0.091 2008-09
## 33  4.6 8.3       -9.0    0.014    0.127   0.236  0.494   0.357 2010-11
## 34  3.9 0.3        9.9    0.094    0.206   0.149  0.505   0.033 2011-12
## 35  0.0 0.0      -74.1    0.000    0.000   0.170  0.258   0.000 2014-15
## 36 10.5 2.0       -2.1    0.102    0.271   0.247  0.590   0.108 2015-16
## 37  4.3 1.3      -10.2    0.066    0.139   0.145  0.499   0.087 2003-04
## 38  1.6 0.6       16.3    0.033    0.136   0.227  0.579   0.100 2019-20
## 39  2.3 0.7       -9.9    0.040    0.110   0.187  0.572   0.068 2006-07
## 40  4.5 1.1      -13.6    0.025    0.120   0.219  0.475   0.060 2018-19
## 41  1.0 1.4      -11.8    0.008    0.069   0.241  0.477   0.178 2017-18
## 42  0.9 0.3       -0.2    0.056    0.115   0.145  0.402   0.062 1997-98
## 43  0.9 1.3       -2.7    0.009    0.126   0.187  0.412   0.275 2012-13
## 44  5.8 1.5       -8.2    0.072    0.188   0.179  0.506   0.096 2012-13
```

```
## 45  3.2 1.3        9.9    0.018     0.086    0.161  0.593    0.060 2019-20
## 46  1.4 0.1       15.9    0.175     0.165    0.227  0.410    0.031 1998-99
## 47  8.1 1.7       -9.1    0.082     0.207    0.173  0.493    0.099 2014-15
## 48  5.5 2.2       -6.5    0.030     0.146    0.183  0.553    0.116 2019-20
## 49  3.8 0.8        1.3    0.110     0.146    0.163  0.574    0.078 2010-11
## 50  3.7 3.8       -0.3    0.047     0.147    0.164  0.472    0.260 2006-07
## 51  1.8 0.2      -11.8    0.023     0.276    0.107  0.231    0.037 2006-07
## 52  0.8 0.1        7.3    0.167     0.188    0.161  0.500    0.056 1997-98
## 53  2.8 2.8        2.7    0.016     0.099    0.223  0.551    0.159 1996-97
## 54  3.5 2.2       -2.8    0.044     0.084    0.224  0.491    0.114 1997-98
## 55  1.9 0.5       -3.8    0.068     0.169    0.184  0.395    0.092 1999-00
## 56  2.8 3.3       -2.0    0.026     0.078    0.182  0.554    0.179 2018-19
## 57  4.9 0.8       -2.9    0.095     0.194    0.212  0.577    0.071 2016-17
## 58  1.1 0.6       -3.8    0.013     0.108    0.164  0.546    0.104 2002-03
## 59  3.4 0.8       -2.6    0.107     0.154    0.179  0.512    0.090 2012-13
## 60  1.6 1.6        0.9    0.044     0.090    0.182  0.596    0.187 1996-97
## 61  3.7 0.7       -0.9    0.048     0.203    0.183  0.517    0.072 2015-16
## 62  4.0 2.0       -1.2    0.011     0.126    0.210  0.535    0.116 2019-20
## 63  1.6 3.7       -6.5    0.018     0.073    0.202  0.474    0.297 2016-17
## 64  0.7 0.4        3.2    0.056     0.080    0.217  0.431    0.116 1996-97
## 65  1.0 1.7       -1.2    0.026     0.053    0.124  0.231    0.192 1999-00
## 66  5.3 0.6       -2.7    0.105     0.166    0.133  0.588    0.038 2015-16
##                   Player
## 1         Aaron Harrison
## 2         Alex Poythress
## 3        Andrew Harrison
## 4          Anthony Davis
## 5         Antoine Walker
## 6         Archie Goodwin
## 7            Bam Adebayo
## 8         Brandon Knight
## 9            Chuck Hayes
## 10        Dakari Johnson
## 11          Daniel Orton
## 12         Darius Miller
## 13          De'Aaron Fox
## 14       DeAndre Liggins
## 15      DeMarcus Cousins
## 16        Derek Anderson
## 17          Devin Booker
## 18            Doron Lamb
## 19           Enes Kanter
## 20          Eric Bledsoe
## 21          Erik Daniels
## 22          Gerald Fitch
## 23        Hamidou Diallo
## 24       Isaac Humphries
## 25       Jamaal Magloire
## 26        Jamal Mashburn
## 27          Jamal Murray
## 28           James Young
## 29     Jarred Vanderbilt
## 30         Jeff Sheppard
## 31           Jodie Meeks
```

```
## 32            Joe Crawford
## 33               John Wall
## 34        Josh Harrellson
## 35          Julius Randle
## 36     Karl-Anthony Towns
## 37            Keith Bogans
## 38          Keldon Johnson
## 39       Kelenna Azubuike
## 40          Kevin Knox II
## 41              Malik Monk
## 42               Mark Pope
## 43          Marquis Teague
## 44  Michael Kidd-Gilchrist
## 45           Mychal Mulder
## 46           Nazr Mohammed
## 47            Nerlens Noel
## 48         P.J. Washington
## 49       Patrick Patterson
## 50             Rajon Rondo
## 51         Randolph Morris
## 52           Reggie Hanson
## 53             Rex Chapman
## 54              Ron Mercer
## 55           Scott Padgett
## 56 Shai Gilgeous-Alexander
## 57         Skal Labissiere
## 58         Tayshaun Prince
## 59          Terrence Jones
## 60               Tony Delk
## 61              Trey Lyles
## 62             Tyler Herro
## 63              Tyler Ulis
## 64          Walter McCarty
## 65            Wayne Turner
## 66     Willie Cauley-Stein
```

```
teams_from_college_1 = table(players_from_best_uni_1$team_abbreviation)
mean(players_from_best_uni_1$player_height)
```

```
## [1] 200.2367
```

```
mean(players_from_best_uni_1$player_weight)
```

```
## [1] 99.34352
```

```
mean(players_from_best_uni_1$age)
```

```
## [1] 21.68182
```

```
teams_from_college_1
```

```
##
## ATL BOS CHA CHH CHI CLE DEN DET GSW HOU IND LAC LAL MEM MIA MIN NOH NYK OKC ORL
##   2   6   4   2   1   1   2   2   2   3   1   2   1   1   4   1   2   5   2   4
## PHI PHX SAC SAS UTA WAS
##   4   4   5   1   3   1
```

Based on the combined subsetted data the best program is: - This data has the combination of NBA player of the week and all NBA players.

```
# create a frequency table to understand the university with the highest percentage of student athletes

number_of_players = table(combined_unique_player_set$`Pre-draft Team`)
number_of_players = as.data.frame(number_of_players)
number_of_players = number_of_players%>%
  mutate(per_per_school = (number_of_players$Freq/nrow(number_of_players))*100)
x = max(number_of_players$per_per_school)
x
```

```
## [1] 7.246377
```

```
z = (number_of_players$per_per_school == x)
number_of_players$Var1[z]
```

```
## [1] Kentucky
## 138 Levels: Alabama Alief Elsik High School (Texas) ... Zalgiris (Lithuania)
```

```
#statistical analysis
#college decided based on above code chucks observation
players_from_best_uni=sqldf("
  SELECT *
  FROM   combined_unique_player_set
  WHERE `Pre-draft Team` == 'Kentucky'
  ")
players_from_best_uni
```

```
##      X1        player_name team_abbreviation age player_height player_weight
## 1  7556      Anthony Davis               NOH  20        208.28      99.79024
## 2   159     Antoine Walker               BOS  20        205.74     101.60461
## 3  9616        Bam Adebayo               MIA  20        208.28     115.66596
## 4  6456   DeMarcus Cousins               SAC  20        210.82     122.46984
## 5   711     Derek Anderson               CLE  23        195.58      88.45044
## 6  1817    Jamaal Magloire               CHH  23        208.28     117.93392
## 7    65     Jamal Mashburn               MIA  24        203.20     113.39800
## 8  6285          John Wall               WAS  20        193.04      88.45044
## 9  8781 Karl-Anthony Towns               MIN  20        213.36     110.67645
## 10 4696        Rajon Rondo               BOS  21        185.42      77.56423
##     college country draft_year draft_round draft_number gp  pts  reb ast
## 1  Kentucky     USA       2012           1            1 64 13.5  8.2 1.0
## 2  Kentucky     USA       1996           1            6 82 17.5  9.0 3.2
## 3  Kentucky     USA       2017           1           14 69  6.9  5.5 1.5
## 4  Kentucky     USA       2010           1            5 81 14.1  8.6 2.5
## 5  Kentucky     USA       1997           1           13 66 11.7  2.8 3.4
```

```
## 6  Kentucky  Canada       2000           1           19 74  4.6  4.0 0.4
## 7  Kentucky     USA       1993           1            4 69 11.9  4.3 2.9
## 8  Kentucky     USA       2010           1            1 69 16.4  4.6 8.3
## 9  Kentucky     USA       2015           1            1 82 18.3 10.5 2.0
## 10 Kentucky     USA       2006           1           21 78  6.4  3.7 3.8
##    net_rating oreb_pct dreb_pct usg_pct ts_pct ast_pct  season
## 1        -5.5    0.106    0.236   0.215  0.559   0.060 2012-13
## 2        -9.3    0.103    0.189   0.251  0.474   0.150 1996-97
## 3        -0.6    0.086    0.194   0.157  0.570   0.115 2017-18
## 4        -7.5    0.105    0.247   0.272  0.484   0.149 2010-11
## 5         3.9    0.038    0.091   0.215  0.531   0.219 1997-98
## 6         3.2    0.111    0.195   0.164  0.506   0.042 2000-01
## 7         1.4    0.037    0.121   0.207  0.487   0.158 1996-97
## 8        -9.0    0.014    0.127   0.236  0.494   0.357 2010-11
## 9        -2.1    0.102    0.271   0.247  0.590   0.108 2015-16
## 10       -0.3    0.047    0.147   0.164  0.472   0.260 2006-07
##                Player                   Team Conference       Date Position
## 1       Anthony Davis     Los Angeles Lakers       West 2019-12-09       PF
## 2      Antoine Walker         Boston Celtics       East 2003-01-19        F
## 3         Bam Adebayo             Miami Heat       East 2019-12-16        C
## 4     DeMarcus Cousins   New Orleans Pelicans       West 2017-10-30        C
## 5      Derek Anderson      San Antonio Spurs       <NA> 2001-03-11        G
## 6     Jamaal Magloire    New Orleans Hornets       East 2004-04-12        C
## 7      Jamal Mashburn       Dallas Mavericks       <NA> 1994-12-11       SF
## 8           John Wall     Washington Wizards       East 2017-03-13       PG
## 9  Karl-Anthony Towns Minnesota Timberwolves       West 2019-10-28        C
## 10        Rajon Rondo         Boston Celtics       East 2010-11-01       PG
##    Height Weight Draft Year Seasons in league Season short Pre-draft Team
## 1    6’10    253       2012                 7         2020       Kentucky
## 2     6’9    265       1996                 6         2003       Kentucky
## 3    6’10    255       2017                 2         2020       Kentucky
## 4    6’11    270       2010                 7         2018       Kentucky
## 5     6’5    194       1997                 3         2001       Kentucky
## 6    6’11    259       2000                 3         2004       Kentucky
## 7     6’8    247       1993                 1         1995       Kentucky
## 8     6’4    210       2010                 6         2017       Kentucky
## 9     7’0    248       2015                 4         2020       Kentucky
## 10    6’1    186       2006                 4         2011       Kentucky
##    Real_value Height CM Weight KG Last Season new_age new_season  wh_ratio
## 1         0.5       208       114           1      26  2019-2020 0.4791158
## 2         0.5       206       120           0      26  2002-2003 0.4938496
## 3         0.5       208       115           1      22  2019-2020 0.5553388
## 4         0.5       211       122           0      27  2017-2018 0.5809214
## 5         1.0       196        88           0      26  2000-2001 0.4522469
## 6         0.5       211       117           0      25  2003-2004 0.5662278
## 7         1.0       203       112           0      22  1994-1995 0.5580610
## 8         0.5       193        95           0      26  2016-2017 0.4581975
## 9         0.5       213       112           1      24  2019-2020 0.5187310
## 10        0.5       185        84           0      25  2010-2011 0.4183164
```

```
teams_from_college = table(players_from_best_uni$Team)
conference_from_college = table(players_from_best_uni$Conference)
best_pos_from_college = table(players_from_best_uni$Position)
teams_from_college
```

```
##
##          Boston Celtics       Dallas Mavericks     Los Angeles Lakers
##                       2                      1                      1
##              Miami Heat Minnesota Timberwolves    New Orleans Hornets
##                       1                      1                      1
##    New Orleans Pelicans      San Antonio Spurs     Washington Wizards
##                       1                      1                      1
```

```
conference_from_college
```

```
##
## East West
##    5    3
```

```
best_pos_from_college
```

```
##
##  C  F  G PF PG SF
##  4  1  1  1  2  1
```

Based on Player of the week data only:

```
# create a frequency table to understand the university with the highest percentage of student athletes

distinct_players = distinct(NBA_player_of_the_week, Player, .keep_all = TRUE)
number_of_players_2 = table(distinct_players$`Pre-draft Team`)
number_of_players_2 = as.data.frame(number_of_players_2)
number_of_players_2 = number_of_players_2%>%
  mutate(per_per_school = (number_of_players_2$Freq/nrow(number_of_players_2))*100)
x2 = max(number_of_players_2$per_per_school)
z2 = (number_of_players_2$per_per_school == x2)
number_of_players_2$Var1[z2]
```

```
## [1] Kentucky
## 169 Levels: Alabama Albany State (GA) ... Zalgiris (Lithuania)
```

```
# statistical analysis
#college decided based on above code chucks observation
alumunia_kent = NBA_player_of_the_week %>%
  filter(`Pre-draft Team` == 'Kentucky')
alumunia_kent = distinct(alumunia_kent, Player, .keep_all = TRUE)
table(alumunia_kent$Position)
```

```
##
##  C  F  G GF PF PG SF
##  4  1  1  1  1  2  1
```

```
mean(alumunia_kent$`Height CM`)
```

```
## [1] 202.7273
```

```
mean(alumunia_kent$Weight)
```

```
## [1] 236.0909
```

```
mean(alumunia_kent$Age)
```

```
## [1] 25
```

Based of the above three methods:

```
number_of_players$Var1[z]
```

```
## [1] Kentucky
## 138 Levels: Alabama Alief Elsik High School (Texas) ... Zalgiris (Lithuania)
```

```
number_of_players_1$Var1[z1]
```

```
## [1] Kentucky
## 307 Levels: Alabama Alabama A&M Alabama Huntsville ... Yonsei (KOR)
```

```
number_of_players_2$Var1[z2]
```

```
## [1] Kentucky
## 169 Levels: Alabama Albany State (GA) ... Zalgiris (Lithuania)
```

Based on the above data classification we can thereby come to the conclusion that the best university in the program is that of Kentucky.

Parameters for success while in the Kentucky program: Height : 200.2367-202.7273 Weight : 236.0909pounds or 99.34352kg

Average age at which athletes from Kentucky get drafted: 21.68182

Most successful position to make it the the NBA from the program: Center

Based on previous data the NBA team you are most likely to go to: Boston Celtics

Based on previous data the conference you are most likely to play in: East

Which positions are most likely to be NBA Player of the Year?

First, we begin with using our merged data set that contains the players of the week with the duplicates removed. From here, we will start with counting the frequency of the position for player of the week and seeing which position appears the most. This would give us an indication for which positions are often the most influential, although we have to consider that part of this comes down to the skill of the player in question.

```
#combined_unique_player_set

positions <- sqldf("
      select position, count(position) as frequency
      from combined_unique_player_set
      group by position
      order by count(position)  desc
      ")
head(positions,20)
```

```
##    Position frequency
## 1         G        44
## 2         C        36
## 3        PG        29
## 4        PF        29
## 5        SG        27
## 6         F        26
## 7        SF        25
## 8        FC        20
## 9        GF        12
## 10      G-F         2
## 11      F-C         1
```

```
toppositions = head(positions,6)
```

Now for a reader with limited knowledge of basketball, the positions as following above go G = Guard, C = Center, PG = Point Guard, PF = power forward, SG = Shooting Guard. F = Forward, SF = Small Forward, FC = forward center, GF = Guard Forward, G-F = Guard Forward. Now this presents us with an issue, because we have to consider that there is some overlap within the positions themselves. On looking at the positions printed out we see that the most prominent position in terms of Players of the Week is the Guard. However this information is insufficient to make a decision on which position would ideally win player of the year. Therefore, we will focus on which positions have the highest frequency of players in terms of assists and points scored based on the data subsets created earlier.

```
assist_positions <- sqldf("
      select position, count(position) as frequency
      from top_50_assisters
      group by position
      order by count(position)  desc
      ")
assist_positions
```

```
##   Position frequency
## 1       PG        27
## 2        G        18
## 3       SF         2
## 4       SG         1
## 5       PF         1
## 6        F         1
```

We are only looking at the top 50 assisters and we notice that the top 6 positions for this list are: PG = Point Guard, G = Guard, SF = Small Forward, SG = Shooting Guard, PF = power forward, F = Forward.

```r
score_positions <- sqldf("
      select position, count(position) as frequency
      from top_50_scorers
      group by position
      order by count(position)  desc
      ")
score_positions
```

```
##    Position frequency
## 1        SF         9
## 2         C         9
## 3        PG         6
## 4        PF         6
## 5         G         6
## 6        SG         5
## 7         F         5
## 8        FC         2
## 9       G-F         1
## 10      F-C         1
```

We are only looking at the top 50 scorers and we notice that the top 6 positions for this list are: PG = Point Guard, G = Guard, SF = Small Forward, SG = Shooting Guard, PF = power forward, C = Center.

We will now combine the two frequency tables obtained for assists and scorers. To do this we will combine the tables by keeping Positions as our comparison variable to create a new table of the common positions between the above mentioned tables(i.e. assists and scorers). We will then combine our frequency and sort this variable to see which position has the highest count.

```r
first_combine = inner_join(score_positions,assist_positions,by = 'Position')
first_combine=first_combine%>%
  mutate(freq = frequency.x+frequency.y)%>%
  select(Position,freq)
first_combine <- sqldf("
      select Position, freq
      from first_combine
      order by freq  desc
      ")
first_combine
```

```
##   Position freq
## 1       PG   33
## 2        G   24
## 3       SF   11
## 4       PF    7
## 5       SG    6
## 6        F    6
```

We will then combine our new table with that created to look at the frequency of the position for player of the week and seeing which position appears the most.

```r
second_combine = inner_join(first_combine,toppositions,by = 'Position')
second_combine=second_combine%>%
```

```
  mutate(com_freq = freq+frequency)
second_combine <- sqldf("
      select *
      from second_combine
      order by com_freq  desc
      ")
second_combine
```

```
##   Position freq frequency com_freq
## 1        G   24        44       68
## 2       PG   33        29       62
## 3       PF    7        29       36
## 4       SG    6        27       33
## 5        F    6        26       32
```

Based on the above information we come to the conclusion that the best 5 positions are: PG = Point Guard, G = Guard, SF = Small Forward, SG = Shooting Guard, PF = power forward. However, it is most likely to win the player of the year if you are either a Point Guard and Guard since they have the highest frequency count across both the tables.