NBA Data Analysis - Skills, Salaries, and Colleges

Jong Ha Lee, Diana Ly, Stacy Chang, Laura Kim

December 13, 2015

Introduction

Almost at the end of every NBA season, the sports media mainly revolves around salaries and new transfers for seasoned NBA players, who continuously record millons of dollar deals as time passes on. However, are the players really worth the amount of money clubs pay for? How can we measure the "worth" of a player? Although there are many aspects, we believed that one of the main ways to measure a player's "worth" is through analyzing their skill attributes most applicable to their position. Our first part of the analysis focuses on correlating most important skill of each position, to the player's salaries.

An equally important aspect of NBA is its rookies, and the colleges that breed these potentially legend NBA players in the future. Thus, we thought it would be interesting to analyze which colleges produce the most number of NBA players currently present, and create a map that displays all the colleges of current NBA players.

Thus, our research question can mainly be detailed in two parts:
1. **In the 2014-2015 season, do the skills of a player correlate to his salary?**
2. **Which college(s) produced the most NBA players in the 2014-2015 season?**

Part 1: Do the skills of a player correlate to his salary?

Data Extraction and Cleaning

Salaries Extraction

We downloaded our salaries from the ESPN Website, where it displayed a total of 480 player's salaries in the 2014-15 season, in a span of 11 pages. Since the website did not provide a specific data frame format (such as csv), we had to use the R Package rvest, which allows us to download data from html websites. We went through the 11 pages in the ESPN website, and produced a raw csv file raw_salaries.csv which will be used in the future for further data cleaning and analysis. All the outlined process was done in download_salaries.r file.

Code we used to download data from ESPN website:

```
for (i in 1:11) {
  page_url <- paste0(webpage, i)
  salary_on_page <- read_html(page_url) %>%
                    html_node(".tablehead") %>%
                    html_table()
  salaries <- rbind(salaries, salary_on_page)
}
```

The Raw CSV produced:

```
##   X X1                 X2                X3          X4
## 1 1 RK               NAME              TEAM      SALARY
## 2 2  1    Kobe Bryant, SF Los Angeles Lakers $23,500,000
## 3 3  2    Joe Johnson, SF      Brooklyn Nets $23,180,790
## 4 4  3 Carmelo Anthony, SF    New York Knicks $22,458,401
## 5 5  4   Dwight Howard, C    Houston Rockets $21,436,271
## 6 6  5     Chris Bosh, PF         Miami Heat $20,644,400
```

Salaries Cleaning

After we downloaded our raw data raw_salaries.csv, data cleaning was required. First, we made sure to only have rows with player salary data, since the rawdata contained a row of headers (RK, NAME, TEAM, SALARY) for every page of ESPN website we went through.

Next, we wanted to separate the player's name and position, which were currently together in the data frame, separated by a comma and a space. `strsplit` function was used, and we created a new column on the data frame, storing the positions.

Lastly, we wanted to convert the `SALARY` column to a numeric, so that further data analysis can be made. This included getting rid of the $ signs and comma's through `gsub` function, and converting the resulting column into a numeric. This cleaned data was exported as salaries.csv in the data folder.

Cleaned data:

```
##   X          Player                Team   Salary Position
## 1 2     Kobe Bryant  Los Angeles Lakers 23500000       SF
## 2 3     Joe Johnson       Brooklyn Nets 23180790       SF
## 3 4 Carmelo Anthony     New York Knicks 22458401       SF
## 4 5   Dwight Howard     Houston Rockets 21436271        C
## 5 6      Chris Bosh          Miami Heat 20644400       PF
## 6 7    LeBron James Cleveland Cavaliers 20644400       SF
```

Roster and Player Statistics: Data Extraction and Cleaning

Our main source of data for both roster (which included the player's name, position, college, and other basic information) came from basketball-reference.com. However, a problem was that the data for all 30 teams could only be downloaded via clicking, without producing a link that we can use in R to download or read.csv the data. Thus, we had to take a different approach, as outlined:
1. Download the roster and player statistics (Totals) of each team by going into each team's website and clicking on "Export", which downloaded the csv files to our computer.
2. Upload the data files into our github repository so anyone can access the raw data via a URL link.

Thus, though in our rawdata directory the csv files for the teams exist, we never directly used them, and instead used github links to `read.csv` the files.

In order to avoid copy-and-pasting to `read.csv` the roster and player data, and aggregate all 30 teams' data into two files respectively, we used string manipulation and a vector of all team names to `read.csv` the data from github through a for loop. Then, we had to do a little cleaning of data specifically for player statistics data, where it included Team Total Statistics as well which was not necessary for our analysis.

The roster data of all 30 teams in the NBA:

```
##   X No.                 Player Pos   Ht  Wt       Birth.Date Exp
## 1 1  19         Furkan Aldemir  PF 6-10 240    August 9 1991   R
## 2 2   0          Isaiah Canaan  PG  6-0 201      May 21 1991   1
```

```
## 3 3   1 Michael Carter-Williams  PG  6-6 190  October 10 1991   1
## 4 4  33        Robert Covington  SF  6-9 215 December 14 1990   1
## 5 5   0          Brandon Davies  PF 6-10 240     July 25 1991   1
## 6 6   7              Larry Drew  PG  6-2 180     March 5 1990   R
##                                College  Team
## 1                                      76ers
## 2              Murray State University 76ers
## 3                  Syracuse University 76ers
## 4           Tennessee State University 76ers
## 5             Brigham Young University 76ers
## 6 University of California Los Angeles 76ers
```

The player statistics data of all 30 teams in the NBA:

```
##   X Rk                  Player Age  G GS   MP  FG FGA   FG. X3P X3PA  X3P.
## 1 1  1            Nerlens Noel  20 75 71 2311 302 653 0.462   0    0    NA
## 2 2  2        Robert Covington  24 70 49 1956 299 756 0.396 167  446 0.374
## 3 3  3        Luc Mbah a Moute  28 67 61 1916 251 636 0.395  62  202 0.307
## 4 4  4         Hollis Thompson  23 71 23 1776 224 543 0.413 115  287 0.401
## 5 5  5              Henry Sims  24 73 32 1399 238 502 0.474   4   22 0.182
## 6 6  6 Michael Carter-Williams  23 41 38 1391 232 611 0.380  32  125 0.256
##   X2P X2PA  X2P.  eFG.  FT FTA   FT. ORB DRB TRB AST STL BLK TOV  PF PTS
## 1 302  653 0.462 0.462 140 230 0.609 185 426 611 128 133 142 146 208 744
## 2 132  310 0.426 0.506 178 217 0.820  65 251 316 105  97  31 128 189 943
## 3 189  434 0.435 0.443  96 163 0.589  82 246 328 106  81  21  99 104 660
## 4 109  256 0.426 0.518  63  89 0.708  53 145 198  85  57  26  66 141 626
## 5 234  480 0.488 0.478 106 137 0.774 121 238 359  79  39  30  99 135 586
## 6 200  486 0.412 0.406 117 182 0.643  43 211 254 302  60  18 174 103 613
##    Team
## 1 76ers
## 2 76ers
## 3 76ers
## 4 76ers
## 5 76ers
## 6 76ers
```

Aggregating Salary, Roster, and Player Statistics Data

Note: Though the cleaned roster data will not be used in this section of the project, we still described roster data and extraction in the previous section, for coherence of data extraction and cleaning method.

Finally, we aggregated the Player Statistics and Salary data, using `dplyr` package function `left_join`, aggregating the data based on the player name, or the column `Player`. However, unfortunately ESPN did not have the salary data for all players, and resulted in some players in the new aggregated data set `stats_salary` not having a value for salary. We omitted the players without the salary data for a more accurate representation of our analysis in the future.

Another problem was that players get traded mid-season, and thus had multiple statistics, in different teams, for the same player. Thus this resulted in three different rows in the data frame for the same player, and we decided to use only one of the three rows to not have data with duplicate salaries, but outstandingly different player statistics in different teams.

This final data frame with the aggregated data was named `stats_salary_pos`, as displayed below:

```
##    X Rk                  Player Age  G GS   MP  FG FGA   FG. X3P X3PA
## 1  1  1            Nerlens Noel  20 75 71 2311 302 653 0.462   0    0
## 2  2  2        Robert Covington  24 70 49 1956 299 756 0.396 167  446
## 3  6  6 Michael Carter-Williams  23 41 38 1391 232 611 0.380  32  125
## 4  7  7            Jerami Grant  20 65 11 1377 124 352 0.352  49  156
## 5  9  9          JaKarr Sampson  21 74 32 1131 146 346 0.422  31  127
## 6 10 10             Tony Wroten  21 30 15  895 175 434 0.403  37  142
##    X3P. X2P X2PA  X2P.  eFG.  FT FTA   FT. ORB DRB TRB AST STL BLK TOV  PF
## 1    NA 302  653 0.462 0.462 140 230 0.609 185 426 611 128 133 142 146 208
## 2 0.374 132  310 0.426 0.506 178 217 0.820  65 251 316 105  97  31 128 189
## 3 0.256 200  486 0.412 0.406 117 182 0.643  43 211 254 302  60  18 174 103
## 4 0.314  75  196 0.383 0.422 114 193 0.591  49 149 198  79  40  68  85 144
## 5 0.244 115  219 0.525 0.467  63  94 0.670  35 128 163  77  38  26  76 135
## 6 0.261 138  292 0.473 0.446 120 180 0.667  22  64  86 157  48   8 113  72
##   PTS Team.x  Salary Position
## 1 744  76ers 3315120       PF
## 2 943  76ers 1000000       SF
## 3 613  76ers 2300040       PG
## 4 411  76ers  884879       SF
## 5 386  76ers  507336       SG
## 6 507  76ers 1210080       SG
```

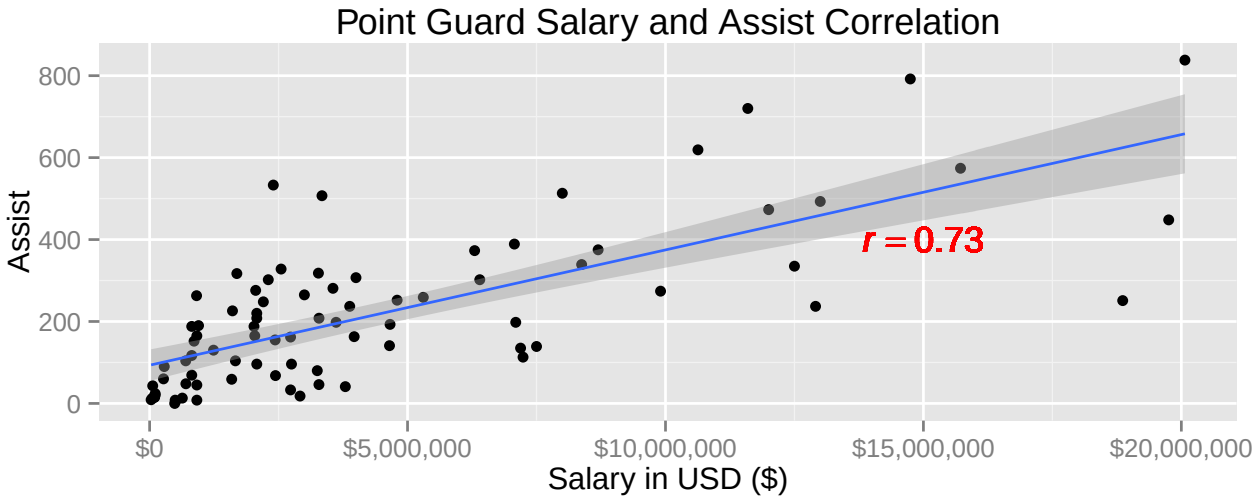
Exploratory Data Analysis: Player Statistics and Roster

After data aggregation and cleaning, we were set to conduct data analysis to answer our research question: *Do the skills of a player correlate to his salary?*

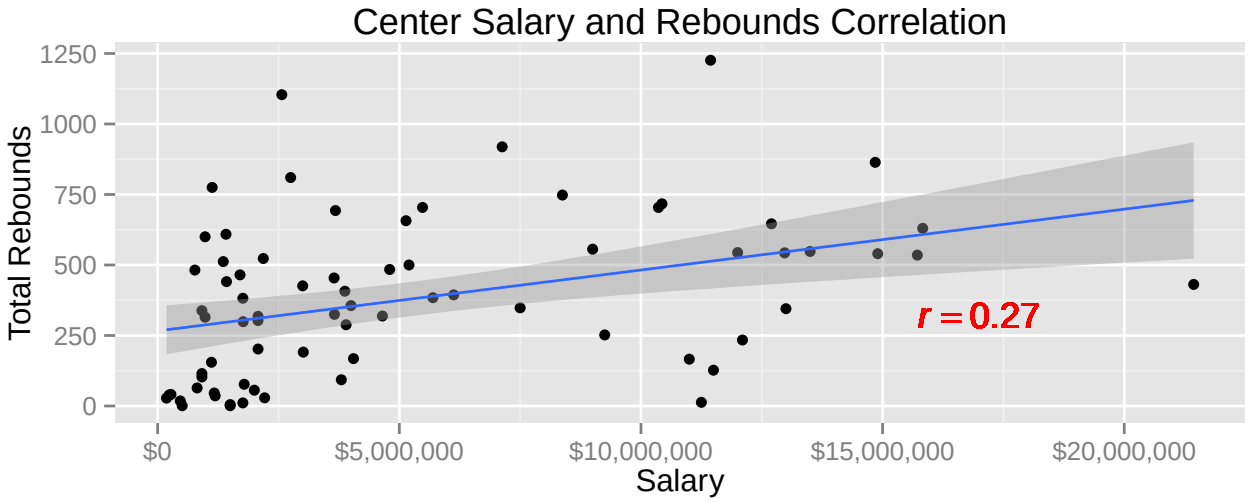
First, we created a function `corr_eqn` which calcualted the correlation coefficient between two input numeric columns. This correlation equation would be used to show our correlation coefficent, and our linear regression line.

Then, we correlated the best skill suited for each position (as outlined below) with salary, creating a scatterplot which included the correlaiton coefficient and the linear regression line: * Point Guard: Assist * Center: Total Rebound * Small Forward: 3 Points Percentage * Power Forward: Total Rebound * Shooting Guard: Field Goal Percentage

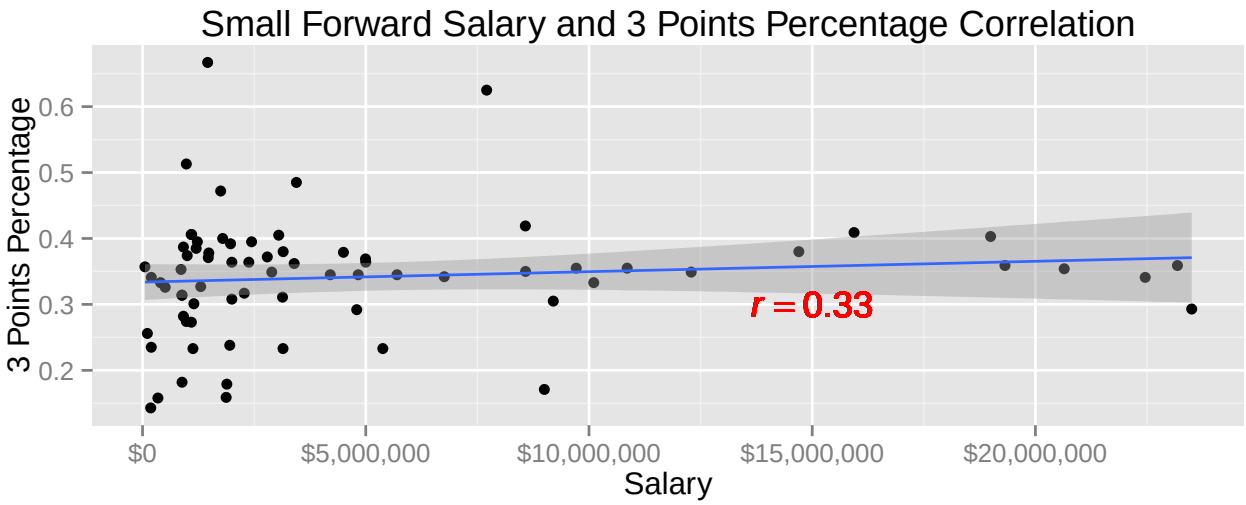
Point Guard Correlation Plot:



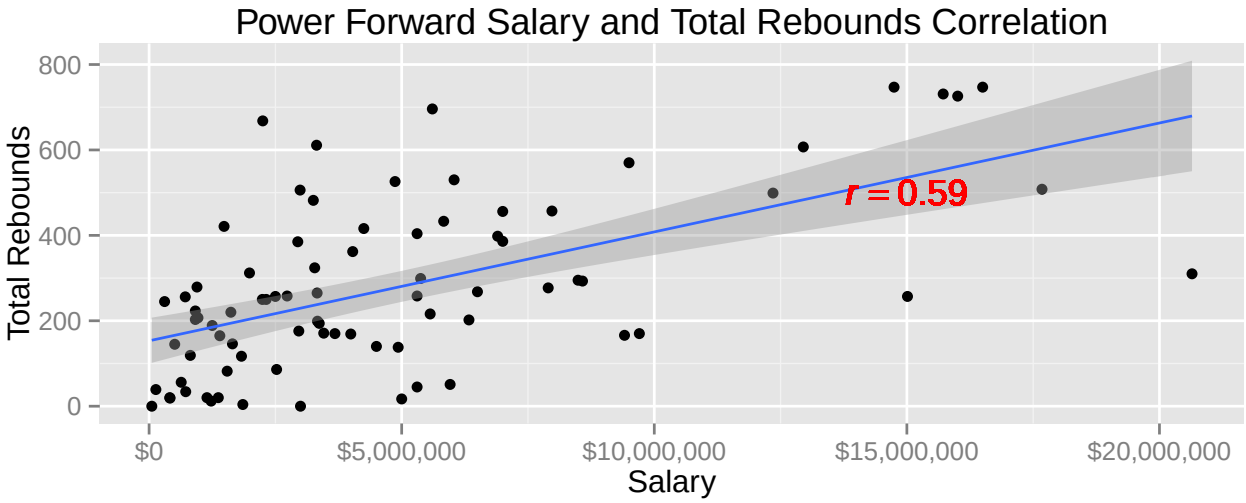
Center Correlation Plot:



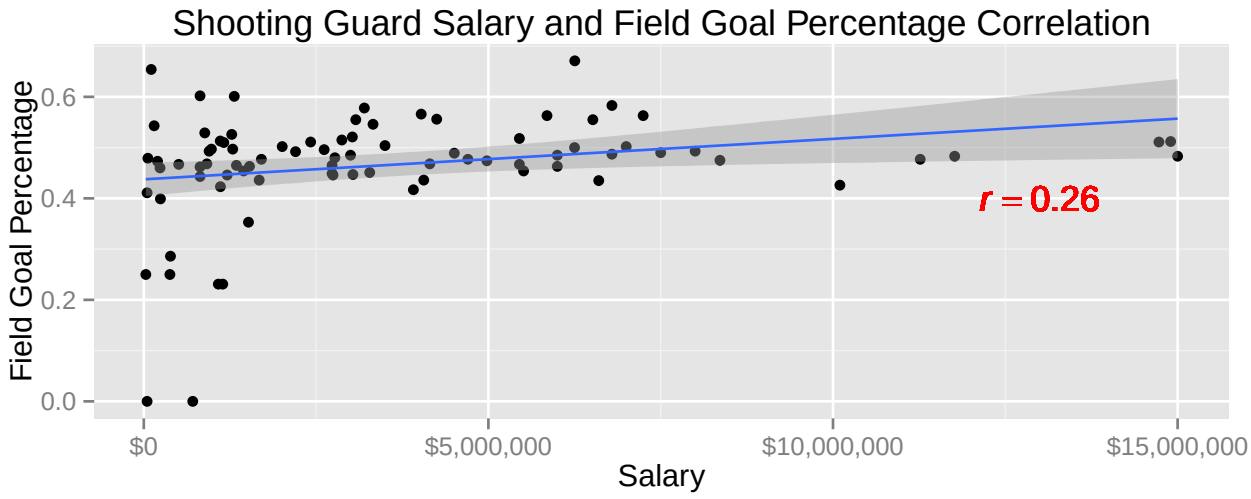
Small Forward Correlation Plot:



Power Forward Correlation Plot:



Shooting Guard Correlation Plot:



General Analysis

These plots indicate mixed results. For example, it seems that point guard salary and the number of total assists are pretty well correlated (r = 0.73), whereas center salary and the total number of rebounds is extremely loosely correlated, with r = 0.27. Also, something noticeable are the effects of certain outliers in either salary (in $), or the skil that is being correlated with salary, differing per position. For example, in the Shooting Guard vs. Field Goal Percentage plot, there are outliers in both field goal percentage, and the salary. By excluding these outliers, the plot could have possibly had better correlation coefficient.

In addition, one must acknowledge that basketball is inherently a team sport. There are cases where the point guard gets many rebounds and scores many points (working as a dual-guard, for example), or when the center has a high field goal percentage because they simply don't shoot much, but make the correct shots that have a high percentage of scoring. In addition, while a player's worth is dependent on the relevant skillset of his position, his value does not only come from such aspect; it may come from his teamwork, leadership, experience, or many other traits he has as a person, and as a player.

Acknowledgeably there are many more aspects to analyze a player, but our data analysis shows an interesting fact: The relevant skillset to a position of a player, and his salary, does not necessarily correlate well with one another.

Part 2: Which college(s) produced the most NBA players in the 2014-2015 season?

Now, we move on to the next part of our project, where we analyze colleges and the number of NBA players each institution has produced. Also, we will produce a map of institutions that have produced NBA players who are present in the 2014-15 season.

Data Cleaning and Extraction

Downloading Data: US Colleges

First, to create our desired map, we had to download US Colleges data so that we can get each institution's geographical data (longitude and latitude). This also posed a slight problem in that there was only a URL to download the zip file that contained multiple csv files. Thus, we downloaded the zip file via URL, and put it in our rawdata directory.

Next, we unzipped the `US_colleges_raw.zip` file, selected the dataset we wanted, and selected the columns we wanted from that data set, which were: Institution name, Longitude, and Latitude. This data set was written as a new csv file, named `US_colleges.csv` in the data directory.

US_Colleges Data (Cleaned):

```
##   X                             College Longitude Latitude
## 1 1  Community College of the Air Force -86.24455 32.40614
## 2 2            Alabama A & M University -86.56850 34.78337
## 3 3 University of Alabama at Birmingham -86.80917 33.50223
## 4 4                  Amridge University -86.17401 32.36261
## 5 5 University of Alabama in Huntsville -86.63842 34.72282
## 6 6            Alabama State University -86.29568 32.36432
```

Merging Data: Roster and US Colleges

Now, we will be using the `roster.csv` data set we cleaned in the previous part of the project, and merge it with US Colleges geographical data.

First, we made sure to not have duplicate of same player in different teams resulting from NBA trades during the season, by using the user-created function `unique_data` (refer to functions.r file for more documentation). Then, after cleaning the names of institutions, we merged the roster of players and the US Colleges geographical data, based on the colleges players attend.

However, a problem arose where US_Colleges data was too overarching, containing multiple branches of colleges in different locations. Thus, we had no choice but to hard-code and research individually which branch of college the players were from, or which branch was the main branch of the college.

Here is an example code:

```
player_colleges$Longitude[pcc == "University of Tennessee"] <-
  US_colleges$Longitude[grep("(The University of Tennessee Knoxville)", usc)]
player_colleges$Latitude[pcc == "University of Tennessee"] <-
  US_colleges$Latitude[grep("(The University of Tennessee Knoxville)", usc)]
```

After the tedious work, we were able to connect the geographical data of the right college branch, with the college in which NBA players played in during their college years. Lastly, for those players who did not play in college, we got rid of `NA`'s and put "No College/University" to prevent future potential problems.

The final merged dataset is displayed below:

```
##   X.1 X No.                  Player Pos   Ht  Wt       Birth.Date Exp
## 1   1 1  19          Furkan Aldemir  PF 6-10 240    August 9 1991   R
## 2   2 2   0           Isaiah Canaan  PG  6-0 201      May 21 1991   1
## 3   3 3   1 Michael Carter-Williams  PG  6-6 190  October 10 1991   1
## 4   4 4  33        Robert Covington  SF  6-9 215 December 14 1990   1
```

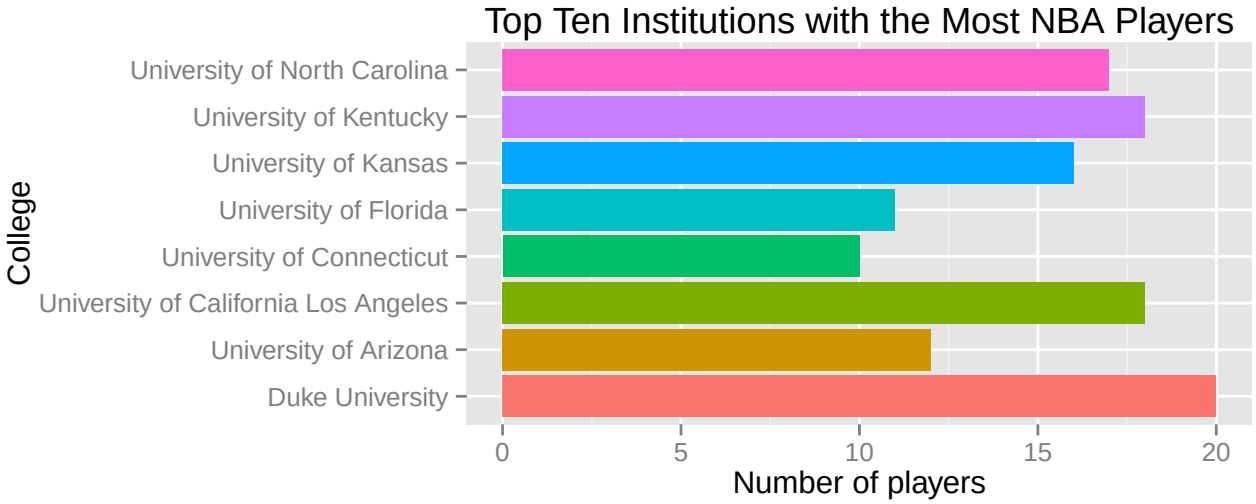
```
## 5   5 5   0          Brandon Davies  PF 6-10 240     July 25 1991   1
## 6   6 6   7              Larry Drew  PG  6-2 180     March 5 1990   R
##                                 College  Team  Longitude Latitude
## 1              No college / university 76ers         NA       NA
## 2              Murray State University 76ers  -88.32349 36.61241
## 3                  Syracuse University 76ers  -76.13674 43.04053
## 4           Tennessee State University 76ers  -86.82937 36.16899
## 5             Brigham Young University 76ers -111.64928 40.25085
## 6 University of California Los Angeles 76ers -118.44390 34.06889
```

Exploratory Analysis

With our data sets, we were able to analyze the top ten colleges that produced the most number of NBA players present in 2014-2015 season, by converting the merged dataset into a table where frequency of college can be counted. We merged this table with the preexisting data frame of `player_colleges`, merging based on College.

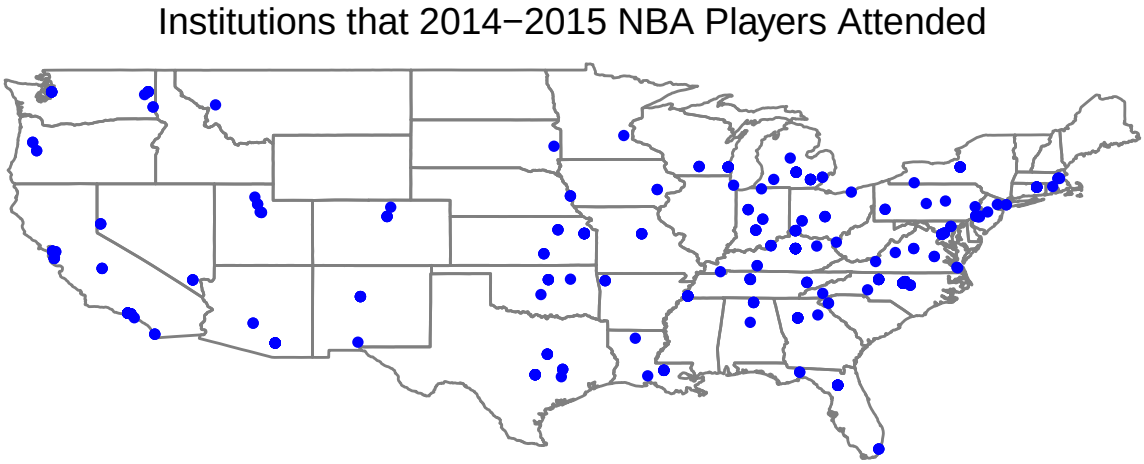
Then, We created a barchart through ggplot, indicating the number of players that graduated from these top ten institutions, for each institution:

The bar plot is displayed below:



Next, we used the R package `maps` to pinpoint the locations of all institutions that produced NBA players present in the 2014-2015 season on a United States map.

The map is displayed below:



General Analysis

As clearly shown, and as expected, the institutions regarded with the best basketball program in the country has produced the most NBA players still present in the 2014-2015 season. Thus, the top ten institutions does not include institutions that have recently began to rise up, upgrading its basketball program, because players still present were once a rookie 10-20 years ago, which implies that these institutions have consistently been producing great NBA players for 10-20 years. However, there are also Universities we don't really hear nowadays as the best, such as University of Florida and University of Connecticut.

The map, on the other hand, shows the diversity of colleges these NBA players come from. However, it is surprising to see that none of the institutions in the Mid-North Region, near North and South Dakota, have produced NBA players. NBA Player production is highly concentrated in the East Coast.

Conclusion

In conclusion, our project tried to answer two questions:
1. **In the 2014-2015 season, do the skills of a player correlate to his salary?**
2. **Which college(s) produced the most NBA players in the 2014-2015 season?**

We answered the first question by aggregating salary, position, and player statistics data, and creating a correlational graph between two variables: skills best identified with the positions, and salary. We were able to show a correlational coefficient, and a regression line for each position, and was able to demonstrate that there are many factors to determining a player's salary, or in other words, "worth" besides the skill most attributed to that player's position.

Then, we answered the second question through a barchart that clearly showed the top ten institutions that produced the most NBA players present in 2014-2015 season, and a map that displayed all the colleges NBA players present in 2014-2015 season are from through points. This part of the project was a little more straightforward, but equally important.

This project helped us gain better knowledge on NBA, NCAA colleges, and basketball in general. We hope that others feel the same as well, and hope to do further data analysis, to explore many other different aspects present in basketball.