NBA Players Data Management and Analysis

Yiziying(Kimmi) Chen

Introduction

- **Statement**:

NBA is becoming the most popular sports league in United States of America. Nowadays, the league pays much more attention to the analysis of player's data, and using the data analysis result to design strategies for winning more games, and probably titles. An example figure in Daryl Morey who introduced the *Moneyball Theory* who transformed Houston Rockets into a very competitive team in the league. In a similar fashion, we are also interested in doing data analysis by exploring the elite figures in this league.

- **Background information:**

a. Link to the data source: https://www.kaggle.com/drgilermo/nba-players-stats
b. The data-set contains aggregate individual statistics for 67 NBA seasons (1950-2017) from basic box-score attributes such as points, assists, rebounds etc., to more advanced money-ball like features such as Value Over Replacement.
c. The data is scraped from https://www.basketball-reference.com
d. Data description:
   This file contains the performance of different players. Some important variables in this data set are:
   1) Year: Season
   2) Player: Player Name
   3) Games: Games Played in the Season
   4) Age: Age of the player
   5) Offensive Rebound: Total number of rebounds player gets during offense. (1974)
   6) Defensive Rebound: Total number of rebounds player gets during defense. (1974)
   7) Points: Total number of points player gets in the season.
   8) Assists: Total number of assists the player has in the season.
   9) Blocks: Total number of blocks the player has in the season. (1974)
   10) Steals: Total number of blocks the player has in the season. (1974)
   11) Turnovers: Total number of turnovers the player has in the season. (1978)
   12) Personal Fouls: Total number of fouls the player has in the season.

Methods

- **Description of the original data files.**

Our original data is retrieved from kaggle.com. The data-set contains aggregate individual statistics for more than 3900 basketball players in 67 NBA seasons, from 1950 to 2017.

- **Description of the guidelines used to validate the data.**

We input the original dataset including 2 separate data files named as "season" and "player"respectively into SAS system. We kept all 7 variables from the original "Players" dataset. From "Seasons_Stats" raw dataset, we picked up 46 variables out of the 52 variables, in which "Player", "Team" and "Position" are categorical.

- **Description of the issues needed to be cleaned and how it will be done.**

1. Missing values. In the file "Players", there were some missing values in player's name. In this case, we just deleted the whole entry. In the file "Season_Stats", if either the player's name or the year is missing, we deleted the whole entry because in the data file "Season_Stats", it is sorted by the year, and an entry in that file without a year is meaningless to us. Furthermore, we

dealt with other variables' missing values, such as College and Birthday, by replacing the blank with "N/A".
2. Ununiformed name format. There is '*' at the end of some players' name. Since we didn't know the meaning of that, we removed '*' from original string using 'COMPRESS'.
3. Wrong calculations. Some values were calculated incorrectly based on the entry's other variables. We solved this by re-calculating the values.

- **Description of additional data preparation that you performed.**

To make the 2 imported datasets more explicit for analyzing, we added labels to some variables which are abbreviations of official terms in NBA games. By browsing the 2 datasets, we discovered that 35 is an appropriate number to include all the characters in including complete players' names. 50 characters will conclude the longest college name. We merge two datasets, "player" and "season" into one, named "nba". Furthermore, we specified the starting season to 1984, the year Michael Jordan had his first debut, because since Michael Jordan has brought basketball sports to another level, currently prominent and well known figures appeared after that year so we think the analysis after this year would make our report much more interesting.

In addition, like we said above, there are some calculation errors present in the dataset, and we would like to fix those errors in variables relating to two categories. The first category relates to variables FG (Field Goals), 2P (2-Point Field Goals) and 3P (3-Point Field Goals) which has a general equation of $Rate * Attempts = Field\ Goals$. In dealing with these problems, we use a difference threshold of 5 between ($Rate \times Attempts$) and ($Field\ Goals$) for detecting erroneous measurements in the dataset, since we believe 5 is the tolerance for these types of calculations. At last, for these erroneous measurements we align rates by calculating $Rate = Field\ Goals \div Attempts$. The second category relates to WS (Winning Share). It turns out that Winning Share is equal to summation of Offensive and Defensive Winning Share ($WS = OWS + DWS$). We set an absolute difference threshold of 0.2 for incorrect measurements in the dataset, and for these incorrect measurements we align Winning Share by calculating $WS = OWS + DWS$.

- **Description of variables to be analyzed including attributes such as name and type.**

1. P_2 (2-Point Field Goal) and P_3 (2-Point Field Goal). Both are numeric variables.
   We analyzed these two variables to get the best 2-point and 3-point shooters after Season 1984.
2. AST (Assists), TRB(Total Rebounds), BLK(Blocks). All are numeric variables.
   We sorted these three variables to get the best assister, rebounder and blocker after Season 1984.
3. College, Birth_State, percentage (P_3/PA_3). Only the percentage is numeric variable.
   We created a table based on these values to get players from Illinois and UIUC.

Results

Check for Players with Missing Height, Weight, or Player Name

Obs	Player	Height	Weight
224	nan	.	.

Data cleaning: From this table we could see there is a player without a name. This player does not contain information, so we drop this player from the table in the data cleaning process.

Print the Data Portion for nba2 (After 1984)

Obs	Player	Year	Position	Age	Team	Game	GS	MP	PER	TS_PCT	ORB_PCT	DRB_PCT	TRB_PCT	AST_PCT
1	A.C. Green	1986	PF	22	LAL	82	1	1542	11.8	56.4%	12.4%	15.5%	14.0%	4.2%
2	A.C. Green	1987	PF	23	LAL	79	72	2240	15.7	59.9%	11.2%	18.8%	15.3%	4.6%
3	A.C. Green	1988	PF	24	LAL	82	64	2636	14.5	58.1%	11.1%	19.1%	15.3%	4.5%
4	A.C. Green	1989	PF	25	LAL	82	82	2510	17.8	59.4%	12.3%	20.0%	16.4%	5.5%
5	A.C. Green	1990	PF	26	LAL	82	82	2709	14.7	54.8%	11.5%	18.4%	15.1%	4.6%

Obs	STL_PCT	BLK_PCT	TOV_PCT	USG_PCT	OWS	DWS	WS	WS_48	BPM	VORP	FG	FGA	FG_PCT	P_3
1	1.5%	1.7%	17.7%	14.7%	1.4	2.0	3.3	0.103	0.3	0.9	209	388	53.9%	1
2	1.5%	2.0%	12.5%	14.7%	4.3	3.3	7.6	0.163	1.7	2.1	316	587	53.8%	0
3	1.6%	1.0%	12.9%	14.7%	4.5	3.4	7.9	0.144	1.0	2.0	322	640	50.3%	0
4	1.8%	1.2%	11.5%	17.0%	5.8	3.5	9.4	0.179	2.2	2.7	401	758	52.9%	4
5	1.2%	1.1%	10.7%	17.1%	4.4	3.3	7.7	0.137	0.4	1.6	385	806	47.8%	13

Obs	PA_3	P_PCT_3	P_2	PA_2	P_PCT_2	eFG_PCT	FT	FTA	FT_PCT	ORB	DRB	TRB	AST	STL	BLK	TOV
1	6	16.7%	208	382	54.5%	54.0%	102	167	61.1%	160	221	381	54	49	49	99
2	5	0.0%	316	582	54.3%	53.8%	220	282	78.0%	210	405	615	84	70	80	102
3	2	0.0%	322	638	50.5%	50.3%	293	379	77.3%	245	465	710	93	87	45	120
4	17	23.5%	397	741	53.6%	53.2%	282	359	78.6%	258	481	739	103	94	55	119
5	46	28.3%	372	760	48.9%	48.6%	278	370	75.1%	262	450	712	90	66	50	116

Obs	PF	PTS	Height	Weight	College	Born_Year	Birth_City	Birth_State
1	229	521	203	106	N/A	1960	N/A	N/A
2	171	852	203	106	N/A	1960	N/A	N/A
3	204	937	203	106	N/A	1960	N/A	N/A
4	172	1088	203	106	N/A	1960	N/A	N/A
5	207	1061	203	106	N/A	1960	N/A	N/A

This part shows the first five observation of the merged dataset which has been subset.

Sample Number, Means and Spreads for nba2 (After 1984)

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Year		17443	2001.87	9.5673649	1984.00	2017.00
Age		17443	26.9050622	4.0493514	18.0000000	44.0000000
Game		17443	49.1378203	26.6471793	1.0000000	85.0000000
GS	Game Started	17443	23.4608152	28.5693241	0	83.0000000
MP	Minutes Played	17443	1148.57	922.6861222	0	3533.00
PER	Player Efficiency Rating	17438	12.3746473	6.2859464	-90.6000000	129.1000000
TS_PCT		17367	0.5037853	0.0958618	0	1.1360000
ORB_PCT		17438	0.0610395	0.0503366	0	1.0000000
DRB_PCT		17438	0.1380708	0.0660786	0	1.0000000
TRB_PCT		17438	0.0995750	0.0508639	0	1.0000000
AST_PCT		17438	0.1307068	0.0970334	0	1.0000000
STL_PCT		17438	0.0164731	0.0103813	0	0.2420000
BLK_PCT		17438	0.0147553	0.0185904	0	0.7780000
TOV_PCT		17382	0.1486053	0.0693007	0	1.0000000
USG_PCT	Usage Percentage	17438	0.1886338	0.0550814	0	1.0000000
OWS	Offensive Win Shares	17443	1.2165511	2.0214021	-3.3000000	15.2000000
DWS	Defensive Win Shares	17443	1.1630453	1.2073282	-1.0000000	9.1000000
WS	Win Shares	17443	2.3804506	2.9433760	-2.1000000	21.2000000
WS_48		17438	0.0653266	0.1039601	-2.5190000	2.1230000
BPM	Box Plus/Minus	17443	-2.3821877	4.7874484	-86.7000000	36.2000000
VORP	Value over Replacement Players	17443	0.5461561	1.3322641	-2.6000000	12.4000000
FG	Field point	17443	180.7906323	178.2823401	0	1098.00
FGA	Field point Attempts	17443	393.8372986	374.9940196	0	2279.00
FG_PCT		17356	0.4363861	0.0979577	0	1.0000000
P_3		17443	23.8245141	39.6604871	0	402.0000000
PA_3		17443	67.9305165	105.3261619	0	886.0000000
P_PCT_3		14263	0.2556355	0.1743668	0	1.0000000
P_2		17443	156.9661182	162.9329773	0	1086.00
PA_2		17443	325.9067821	326.3070146	0	2213.00
P_PCT_2		17327	0.4564509	0.1015301	0	1.0000000
eFG_PCT		17356	0.4641447	0.0998287	0	1.5000000
FT		17443	91.7010835	106.6270076	0	833.0000000
FTA		17443	121.6845153	136.0737545	0	972.0000000
FT_PCT		16728	0.7218181	0.1483609	0	1.0000000
ORB	Offensive Rebounds	17443	58.4197099	64.7885964	0	523.0000000
DRB	Defensive Rebounds	17443	142.3883506	141.0354064	0	1007.00
TRB	Total Rebounds	17443	200.8080605	200.1944971	0	1530.00
AST	Number of Assists	17443	108.2149286	136.9237037	0	1164.00
STL	Number of Steals	17443	37.9303446	37.2608134	0	301.0000000
BLK	Number of Blocks	17443	23.7451700	35.9919632	0	456.0000000
TOV	Number of Turnovers	17443	70.2117755	65.1306352	0	464.0000000
PF	Personal Fouls	17443	105.6234019	79.3964516	0	386.0000000
PTS	Points	17443	477.1068624	474.9701248	0	3041.00
Height		17443	200.2768446	9.5953925	160.0000000	231.0000000
Weight		17443	97.1216534	12.6243700	60.0000000	163.0000000
Born_Year		17443	1972.30	11.9972523	1920.00	1997.00

This is the *proc means* printing results of the merged dataset subset from 1984.

Top Five Players in Assist in the Dataset

Obs	Player	_TYPE_	_FREQ_	total_ast
1	John Stockton	1	19	15806
2	Jason Kidd	1	23	13393
3	Mark Jackson	1	21	11930
4	Steve Nash	1	18	10335
5	Andre Miller	1	25	9717

From this table we can find the top five players good in Assisting.

Top Five Players in Three Point in the Dataset

Obs	Player	College	Birth_State	point	attempt	percentage
66	Steve Kerr	University of Arizona	Lebanon	732	1625	45.0%
68	Hubert Davis	University of North Carolina	North Ca	806	1823	44.2%
69	Stephen Curry	Davidson College	Ohio	1917	4379	43.8%
70	Jason Kapono	University of California, Los Angeles	Californ	457	1054	43.4%
71	Steve Novak	Marquette University	Illinois	627	1446	43.4%

From this table we can see the top five players in three-point percentage shooting.

Top Five Players in rebounds in the dataset

Obs	Player	trbsum
1	Tim Duncan	15091
2	Kevin Garnett	14973
3	Karl Malone	14968
4	Kevin Willis	13803
5	Hakeem Olajuwon	13748

Top Five Players in Blocks in the Dataset

Obs	Player	_TYPE_	_FREQ_	total_blk
1	Hakeem Olajuwon	1	18	3830
2	Dikembe Mutombo	1	20	3492
3	Tim Duncan	1	19	3020
4	David Robinson	1	14	2954
5	Patrick Ewing	1	18	2895

From this table we can see the top five players in rebounds and blocks. We can see that Tim Duncan and Hakeem Olajuwon were both ranked top five twice.

Top Five Players in Field Goals in the Dataset

Obs	Player	_TYPE_	_FREQ_	total_fg	total_fga	fg_PCT
17	DeAndre Jordan	1	9	2487	3674	0.67692
40	Clint Capela	1	3	607	989	0.61375
41	Rudy Gobert	1	4	905	1479	0.61190
42	Artis Gilmore	1	7	1850	3036	0.60935
43	Brandan Wright	1	14	1505	2473	0.60857

From this table we could see the top five players in data analysis.

NBA Players who Attended UIUC

Player	Birth_State	College
Brian Cook	Illinois	University of Illinois at Urbana-Champaign
Bruce Douglas	Illinois	University of Illinois at Urbana-Champaign
Derek Harper	Georgia	University of Illinois at Urbana-Champaign
Deron Williams	West Vir	University of Illinois at Urbana-Champaign
Derrick Williams	West Vir	University of Illinois at Urbana-Champaign
James Augustine	Illinois	University of Illinois at Urbana-Champaign
Ken Norman	Illinois	University of Illinois at Urbana-Champaign
Kendall Gill	Illinois	University of Illinois at Urbana-Champaign
Kenny Battle	Illinois	University of Illinois at Urbana-Champaign
Kiwane Garris	Illinois	University of Illinois at Urbana-Champaign
Luther Head	Illinois	University of Illinois at Urbana-Champaign
Marcus Liberty	Illinois	University of Illinois at Urbana-Champaign
Meyers Leonard	Virginia	University of Illinois at Urbana-Champaign
Nick Anderson	Illinois	University of Illinois at Urbana-Champaign
Robert Archibald	United K	University of Illinois at Urbana-Champaign
Roger Powell	Illinois	University of Illinois at Urbana-Champaign
Scott Meents	Illinois	University of Illinois at Urbana-Champaign
Steve Bardo	Kentucky	University of Illinois at Urbana-Champaign

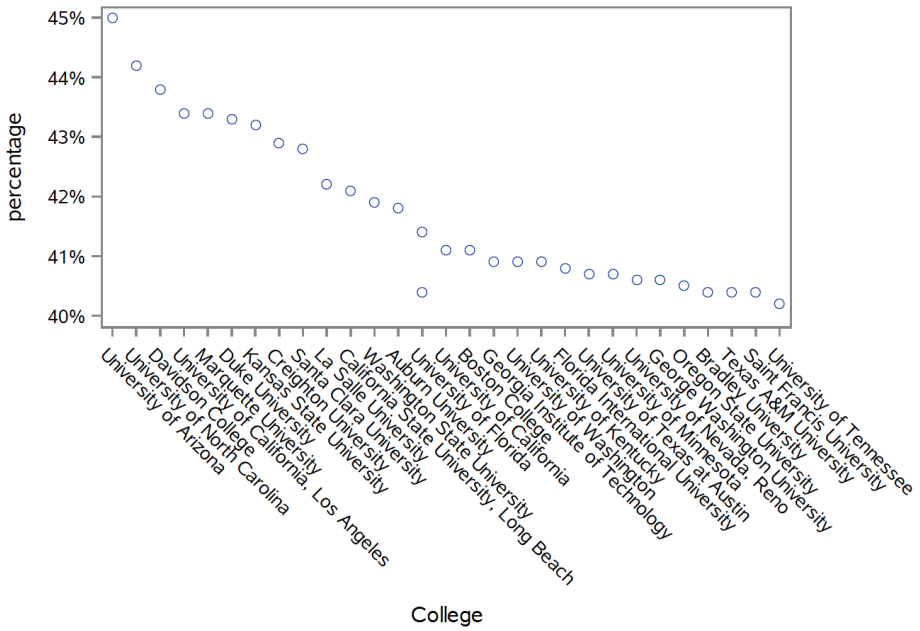
NBA Players Born in Illinois and attended UIUC

Player	Birth_State	College
Brian Cook	Illinois	University of Illinois at Urbana-Champaign
Bruce Douglas	Illinois	University of Illinois at Urbana-Champaign
James Augustine	Illinois	University of Illinois at Urbana-Champaign
Ken Norman	Illinois	University of Illinois at Urbana-Champaign
Kendall Gill	Illinois	University of Illinois at Urbana-Champaign
Kenny Battle	Illinois	University of Illinois at Urbana-Champaign
Kiwane Garris	Illinois	University of Illinois at Urbana-Champaign
Luther Head	Illinois	University of Illinois at Urbana-Champaign
Marcus Liberty	Illinois	University of Illinois at Urbana-Champaign
Nick Anderson	Illinois	University of Illinois at Urbana-Champaign
Roger Powell	Illinois	University of Illinois at Urbana-Champaign
Scott Meents	Illinois	University of Illinois at Urbana-Champaign

The table on the left lists all the NBA players who attended college at University of illinois at Urbana-Champaign, the alumni of U of I.

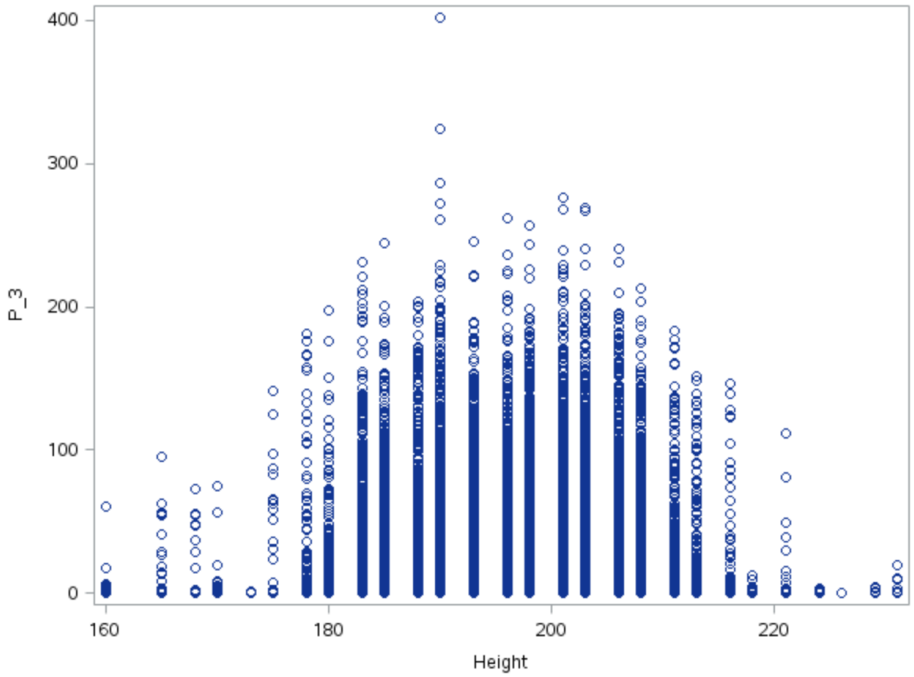
The tale on the right lists all the NBA players who attended college at University of illinois at Urbana-Champaign and came from Illinois.

The Colleges of Top 30 Players with highest 3-Point Shooting Percentage



This plot illustrates the colleges were top 30 NBA players with highest 3-point percentage shooting came from. From the point distribution, we see that the highest 3-point shooting percentage is 45% by a player who attended University of Arizona. All the Colleges listed above have 1 player ranked top 30, expect University of Florida, which has 2 alumni.

The scattered distribution Between Height of each NBA Player and 3-point Made



From the distribution plot, we can interpret a correlation between the Height and 3-point Made. The players between 180 (cm) and 210 (cm) makes the most 3-points. Players within this range usually play position of guards and small forwards, which tend to make more 3-point Attempts.