Syracuse University
School of Information Studies

IST-718 : Big Data Analytics

Project Proposal

NBA PLAYER STATS

Group 9

Ayyappa Sumanth Bhuma
Khushboo Saxena
Kirat Saran
Shreya Balan Nambiar

Objective

Data Analytics has improved the quality of professional basketball. The NBA currently utilizes statistics so enormously, that every team uses data insights to benefit their teams. Legendary players such as Kobe, Jordan, and the current superstars such as Curry, Giannis have their performances recorded in this dataset. Our model aims to use Machine Learning to predict the factors responsible for scouting players. It can also be used for designing winning strategies and avoiding injuries. Factors such as age, average points scored, average number of rebounds grabbed, average number of assists distributed from the years 1996-2021 have been used.

Data Set Description

- Structure of dataset: Record of the performance of the players from 1996-2021
- No of columns: 21
- No of rows: 12305
- Dataset link : https://www.kaggle.com/datasets/justinas/nba-players-data?select=all_seasons.csv
- Interesting/Surprising about data: Younger players are being drafted as years pass. Also, shorter players seem to be better assists.

Data Exploration

Schema

```
root
 |-- player_name: string (nullable = true)
 |-- team_abbreviation: string (nullable = true)
 |-- age: double (nullable = true)
 |-- player_height: double (nullable = true)
 |-- player_weight: double (nullable = true)
 |-- college: string (nullable = true)
 |-- country: string (nullable = true)
 |-- draft_year: string (nullable = true)
 |-- draft_round: string (nullable = true)
 |-- draft_number: string (nullable = true)
 |-- gp: integer (nullable = true)
 |-- pts: double (nullable = true)
 |-- reb: double (nullable = true)
 |-- ast: double (nullable = true)
 |-- net_rating: double (nullable = true)
 |-- oreb_pct: double (nullable = true)
 |-- dreb_pct: double (nullable = true)
 |-- usg_pct: double (nullable = true)
 |-- ts_pct: double (nullable = true)
 |-- ast_pct: double (nullable = true)
 |-- season: string (nullable = true)
```

Our Dataframe

player_name	team_abbreviation	age	player_height	player_weight	college	country	draft_year	draft_round	draft_number	gp
Dennis Rodman	CHI	36.0	198.12	99.79024	Southeastern Okla...	USA	1986	2	27	55
Dwayne Schintzius	LAC	28.0	215.9	117.93392	Florida	USA	1990	1	24	15
Earl Cureton	TOR	39.0	205.74	95.25432	Detroit Mercy	USA	1979	3	58	9
Ed O'Bannon	DAL	24.0	203.2	100.697424	UCLA	USA	1995	1	9	64
Ed Pinckney	MIA	34.0	205.74	108.86208	Villanova	USA	1985	1	10	27
Eddie Johnson	HOU	38.0	200.66	97.52228000000001	Illinois	USA	1981	2	29	52
Eddie Jones	LAL	25.0	198.12	86.18248	Temple	USA	1994	1	10	80
Elden Campbell	LAL	28.0	213.36	113.398	Clemson	USA	1990	1	27	77
Eldridge Recasner	ATL	29.0	193.04	86.18248	Washington	USA	1992	Undrafted	Undrafted	71
Elliot Perry	MIL	28.0	182.88	72.57472	Memphis	USA	1991	2	37	82
Elmer Bennett	DEN	27.0	182.88	77.11064	Notre Dame	USA	1992	2	38	9
Elmore Spencer	SEA	27.0	213.36	122.46983999999999	Nevada-Las Vegas	USA	1992	1	25	1
Emanual Davis	HOU	28.0	195.58	87.996848	Delaware State	USA	Undrafted	Undrafted	Undrafted	13
Ennis Whatley	POR	34.0	190.5	80.28578399999998	Alabama	USA	1983	1	13	3
Eric Leckner	VAN	31.0	210.82	120.20188	Wyoming	USA	1988	1	17	20
Eric Mobley	VAN	27.0	210.82	106.59411999999999	Pittsburgh	USA	1994	1	18	28
Eric Montross	NJN	25.0	213.36	122.46983999999999	North Carolina	USA	1994	1	9	78
Eric Murdock	DEN	29.0	185.42	86.18248	Providence	USA	1991	1	21	12
Eric Piatkowski	LAC	26.0	200.66	97.52228000000001	Nebraska	USA	1994	1	15	65
Eric Snow	SEA	24.0	190.5	90.7184	Michigan State	USA	1995	2	43	67

player_name	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	season
Dennis Rodman	5.7	16.1	3.1	16.1	0.18600000000000003	0.32299999999999995	0.1	0.479	0.113	1996-97
Dwayne Schintzius	2.3	1.5	0.3	12.3	0.078	0.151	0.175	0.43	0.048	1996-97
Earl Cureton	0.8	1.0	0.4	-2.1	0.105	0.102	0.10300000000000001	0.376	0.14800000000000002	1996-97
Ed O'Bannon	3.7	2.3	0.6	-8.7	0.06	0.149	0.16699999999999998	0.39899999999999997	0.077	1996-97
Ed Pinckney	2.4	2.4	0.2	-11.2	0.109	0.179	0.127	0.611	0.04	1996-97
Eddie Johnson	8.2	2.7	1.0	4.1	0.034	0.126	0.22	0.541	0.102	1996-97
Eddie Jones	17.2	4.1	3.4	4.1	0.035	0.091	0.209	0.5589999999999999	0.149	1996-97
Elden Campbell	14.9	8.0	1.6	3.3	0.095	0.183	0.222	0.52	0.087	1996-97
Eldridge Recasner	5.7	1.6	1.3	-0.3	0.036000000000000004	0.076	0.172	0.539	0.141	1996-97
Elliot Perry	6.9	1.5	3.0	-1.2	0.018000000000000002	0.081	0.177	0.557	0.262	1996-97
Elmer Bennett	2.4	0.4	1.2	-7.4	0.0	0.053	0.187	0.47	0.21600000000000003	1996-97
Elmore Spencer	0.0	0.0	0.0	-147.5	0.0	0.0	0.126	0.0	0.0	1996-97
Emanual Davis	5.0	1.7	2.0	6.6	0.011000000000000001	0.098	0.14400000000000002	0.565	0.191	1996-97
Ennis Whatley	1.3	1.0	1.0	-57.4	0.0	0.15	0.115	0.5	0.375	1996-97
Eric Leckner	1.7	1.8	0.3	-3.9	0.044000000000000004	0.258	0.175	0.444	0.067	1996-97
Eric Mobley	2.6	2.1	0.5	-15.4	0.106	0.102	0.158	0.47200000000000003	0.076	1996-97
Eric Montross	4.4	6.6	0.8	-6.6	0.107	0.20600000000000002	0.11199999999999999	0.452	0.05	1996-97
Eric Murdock	3.8	0.9	2.0	-1.4	0.01	0.102	0.188	0.588	0.436	1996-97
Eric Piatkowski	6.0	1.6	0.8	2.4	0.071	0.08800000000000001	0.22	0.579	0.111	1996-97
Eric Snow	3.0	1.0	2.4	-1.2	0.027000000000000003	0.085	0.145	0.515	0.33899999999999997	1996-97

Preliminary Data Exploration

```
nba.summary().show()
```

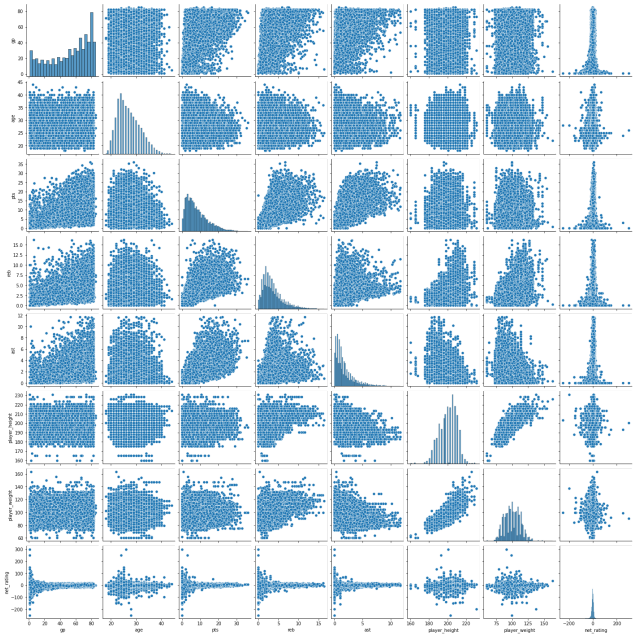
summary	player_name	team_abbreviation	age	player_height	player_weight	college	country	draft_year	draft_round	draft_number
count	12305	12305	12305	12305	12305	12305	12305	12305	12305	12305
mean	null	null	27.08451848841934	200.61160178789982	100.36992618155391	null	null	2003.5106636246403	1.3039577310337953	21.86059034702832
stddev	null	null	4.33586763083955	9.146320529958082	12.477149699854786	null	null	8.756976264650898	0.5071996975631694	16.07449095709673
min	A.C. Green	ATL	18.0	160.02	60.327736		Angola	1963	0	0
25%	null	null	24.0	193.04	90.7184	null	null	1997.0	1.0	9.0
50%	null	null	26.0	200.66	99.79024	null	null	2004.0	1.0	19.0
75%	null	null	30.0	208.28	108.86208	null	null	2010.0	2.0	33.0
max	Zylan Cheatham	WAS	44.0	231.14	163.29312	Yonsei (KOR)	Yugoslavia	Undrafted	Undrafted	Undrafted

summary	gp	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	season
count	12305	12305	12305	12305	12305	12305	12305	12305	12305	12305	12305
mean	51.29053230394149	8.172775294595697	3.5591548151158 2	1.8139861844778555	-2.2557334416903685	0.0544726533929292	0.1410140593254775	0.18489085737505168	0.5110600568874436	0.1313580658268993	null
stddev	25.095908921340524	5.974957082372287	2.4835497662924513	1.7941550768164019	12.673254189413505	0.0435987819930391	0.0627041636557747	0.05338952628276637	0.10116310532933025	0.09448268755642035	null
min	1	0.0	0.0	0.0	-250.0	0.0	0.0	0.0	0.0	0.0	1996-97
25%	31	3.6	1.8	0.6	-6.4	0.021	0.096	0.149	0.48	0.066	null
50%	57	6.7	3.0	1.2	-1.3	0.04099999999999995	0.131	0.18100000000000002	0.524	0.10300000000000001	null
75%	73	11.5	4.7	2.4	3.2	0.084	0.18	0.217	0.561	0.17800000000000002	null
max	85	36.1	16.3	11.7	300.0	1.0	1.0	1.0	1.5	1.0	2021-22

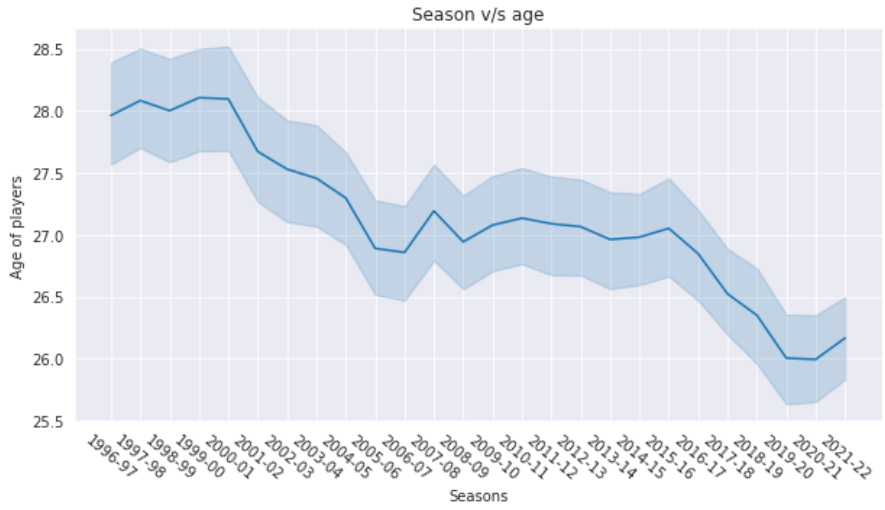
Count of Null and NaN values in our Dataframe

player_name	team_abbreviation	age	player_height	player_weight	college	country	draft_year	draft_round	draft_number	gp	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	season
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Pairplot for the dataset



Age vs Season plot



Proposed Data Exploration Insights

- The NBA data contains variables for players' physical stats, including their height and weight. Our plan is to track these traits' trends over each season. We wish to investigate how the significance of a player's physical attributes altered as the game evolved. One idea is to plot line plots of average values for each season.
- NBA statistics analysis is incomplete without comparing the stats of the best players. In order to compare the top players, we will be looking for metrics that determine how important a player is to the game. Example of a metric is the average percentage of points scored by the player in a game.

- With the help of plots like scatter plots and bar charts, we will be analyzing the distribution of points scored, assists, rebounds, and other game data for players as well as determining the impact of physical attributes and age on the game statistics.
- The majority of the players in NBA data are Americans, however there are also players from other nations. We will conduct a study to contrast the physical characteristics, performance, and game data of players from the United States and around the globe. We seek to distinguish between the physical characteristics and abilities of players from other nations and those from the United States.
- In the NBA, there have been players who began their careers with mediocre stats and rose to the top. So, in order to track players' career development, we shall compare the best players' performances throughout each season.
- We would create a new performance variable based on columns representing no. of games played, average points scored, average no. of rebounds grabbed and average no. of assists distributed throughout the season.

Proposed Predictions

- The net rating, which indicates the net difference in points the team scores (per 100 possessions) depending on whether a player is on the court or not, stands out among the dataset's variables. It is a crucial indicator for determining how valuable a player is to the game. We will create machine learning models to predict a player's net rating based on his personal attributes.
- Basketball is no longer a game where players compete only on the basis of their physical prowess. As a result, the game's emphasis on a player's height, weight, or physique changed. We will create machine learning models to predict the typical physical characteristics of athletes who will play the upcoming season. The goal is to identify the qualities that boost a player's likelihood of being selected in the upcoming draft.
- The percentage of team points a player contributes to is another helpful indicator of their performance. i.e., the player's point total divided by the team's point total. We will develop machine learning models in order to forecast a player's percentage of points in the upcoming game.
- Shooting efficiency which accounts for free throws, 2 and 3 pointers. This is an important indicator to find how valuable an offensive player is.

Model Inference Insights

- Using a linear regression model which predicts net rating of a player, we plan to analyze how each predictor variable affects the final outcome. In addition, we'll look at which characteristics boost a player's rating and which characteristics detract from it.
- We have data from many NBA seasons. We therefore plan to determine the variation in each predictor's impact for several seasons using the models we created.
- We want to see if Age of player has any affect on their shooting efficiency using the model.