4/28/2024

Cortlynd Cox

Data Science

Final Project Presentation

Sample of the Original Data

For this project I used a dataset from Kaggle named “NBA Players.” The description of the dataset is: “Biometric, biographic and basic box score stats from 1996 to 2022 season.” The dataset can be found at

https://www.kaggle.com/datasets/justinas/nba-players-data?resource=download

The dataset needs to be loaded into the ‘sample_data’ folder for my Google Collab program to work.

Looking into this dataset, I found that there were many useful columns that represent the stats of nba players. The rows of the dataset are by player name and year. So any player who played for twenty years would have twenty rows in the dataset representing their stats for each year. The dataset also includes their biometric information, including player weight and height. Here is an example of what the data looks like:

	player_name	team_abbreviation	age	player_height	player_weight	college	country	draft_year	draft_round	draft_number	...	pts	reb	ast	net_rating	oreb_pct	dreb_pct	usg_pct	ts_pct	ast_pct	season
0	Randy Livingston	HOU	22.0	193.04	94.800728	Louisiana State	USA	1996	2	42	...	3.9	1.5	2.4	0.3	0.042	0.071	0.169	0.487	0.248	1996-97
1	Gaylon Nickerson	WAS	28.0	190.50	86.182480	Northwestern Oklahoma	USA	1994	2	34	...	3.8	1.3	0.3	8.9	0.030	0.111	0.174	0.497	0.043	1996-97
2	George Lynch	VAN	26.0	203.20	103.418976	North Carolina	USA	1993	1	12	...	8.3	6.4	1.9	-8.2	0.106	0.185	0.175	0.512	0.125	1996-97
3	George McCloud	LAL	30.0	203.20	102.058200	Florida State	USA	1989	1	7	...	10.2	2.8	1.7	-2.7	0.027	0.111	0.206	0.527	0.125	1996-97
4	George Zidek	DEN	23.0	213.36	119.748288	UCLA	USA	1995	1	22	...	2.8	1.7	0.3	-14.1	0.102	0.169	0.195	0.500	0.064	1996-97
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
12839	Joel Embiid	PHI	29.0	213.36	127.005760	Kansas	Cameroon	2014	1	3	...	33.1	10.2	4.2	8.8	0.057	0.243	0.370	0.655	0.233	2022-23
12840	John Butler Jr.	POR	20.0	213.36	86.182480	Florida State	USA	Undrafted	Undrafted	Undrafted	...	2.4	0.9	0.6	-16.1	0.012	0.065	0.102	0.411	0.066	2022-23
12841	John Collins	ATL	25.0	205.74	102.511792	Wake Forest	USA	2017	1	19	...	13.1	6.5	1.2	-0.2	0.035	0.180	0.168	0.593	0.052	2022-23
12842	Jericho Sims	NYK	24.0	208.28	113.398000	Texas	USA	2021	2	58	...	3.4	4.7	0.5	-6.7	0.117	0.175	0.074	0.780	0.044	2022-23
12843	JaMychal Green	GSW	33.0	205.74	102.965384	Alabama	USA	Undrafted	Undrafted	Undrafted	...	6.4	3.6	0.9	-8.2	0.087	0.164	0.169	0.650	0.094	2022-23

12844 rows × 21 columns

For the variables we have ‘player_name’ which is just the name of the player, ‘team_abbreviation’ which just tells us what team they played for any given season, ‘age’ which tells us how they were that season, ‘player_height’ which tells us how tall they are in centimeters, ‘player_weight’ which tells us how much they weigh in kilograms, ‘college’ which tells us where they played in college, ‘country’ which tells us what country they’re from, ‘draft_year’ which tells us which year they were drafted, ‘draft_round’ which tells us what round of the draft they were drafted in (this could also include being undrafted), ‘draft_number’ which tells us at what pick they were drafted, ‘gp’ which stands for games played and tells us how many games they played in that season, ‘pts’ which tells us how many points they scored on average every game that season, ‘reb’ which tells us how many rebounds they had on average every game that season, ‘ast’ which tells us how many assists they had on average every game that season, ‘net_rating’ which tells us what their plus minus box score is on average for every game that season (plus minus box score is how well the team either outscored or got ourscored by opponents while they were on the court), ‘oreb_pct’ which is how good they were at getting offensive rebounds per possesion, ‘dreb_pct’ which is how good they were at getting defensive rebounds per possession, ‘usg_pct’ which stands for usage percent which the percentage of time a player is used by their team offensively (when they have the ball), ‘ts_pct” which stands for true shooting percentage and tells us how well they shoot weighing for three point shots being worth more points, ‘ast_pct’ which tells us how good a player is at getting assists per possession, and lasly ‘season’ which says what year of that players career the row represents.

Some of my Challenges with the Data

Firstly, when I went to do the part of the writeup that has to do with career rebounds of players, I had to sort the data by unique player names, and then create a dataset that was the average rebounds of any given player over the course of their career. The reason I had to do

this was due to the fact that the data contained many seasons per player, and I wanted to see the data representing their career average. This was a challenge because it required me to learn multiple skills with python, like being able to separate out all unique values in a column, and then being able to use those unique values to create a new dataset that is sorted by those values essentially. I did this sort twice, once for when I was comparing height and rebounding and once when I was comparing weight and rebounding.

When it came to having to clean the data though, I didn't really have to do anything there. I checked to see which columns of the data set had null values, and I found that the only column that had null values was the column representing what college players played at. The reason this one had null values is because not all players in the nba go to college before entering the nba. This could be because they entered the league straight out of high school (notably LeBron James did this), or it could be because they are from another country where it is common for players to play in the European league and then go to the nba. The fact that this column had null values wasn't a problem though, due to the fact that it only would've come up if I had decided to work with the college column, which I did not end up doing.

Explaining What the Figures and Graphs Tell Us

Figure 1:

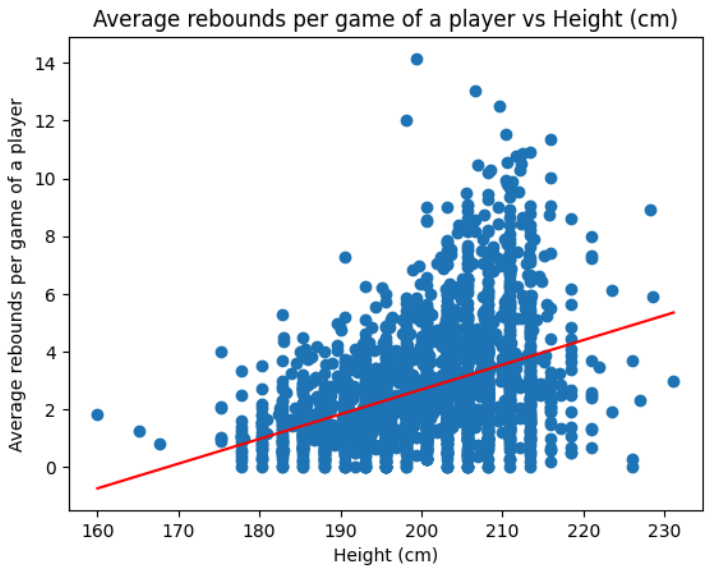


Figure 2:

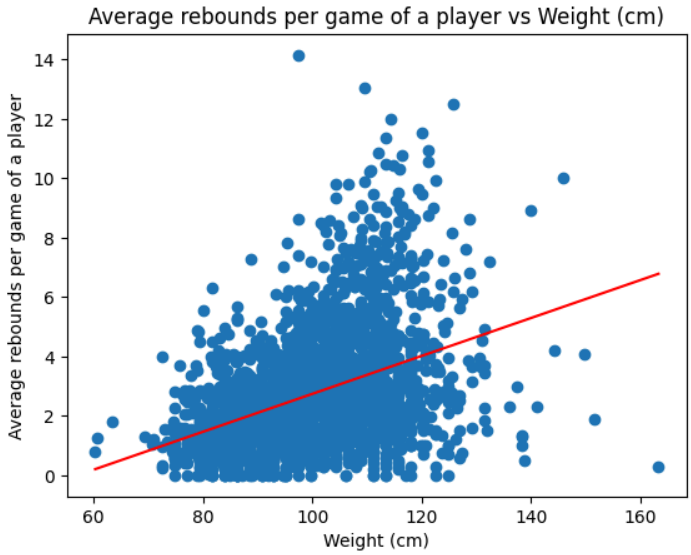


Figure 3:

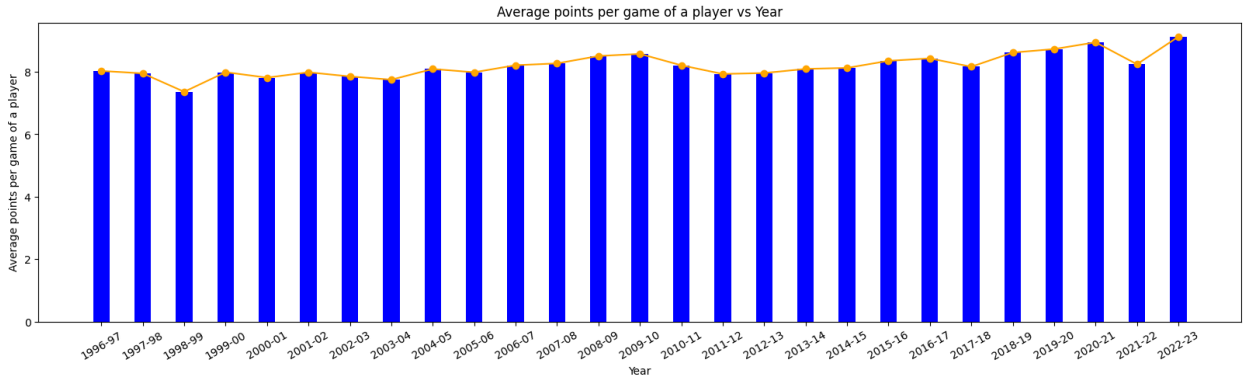
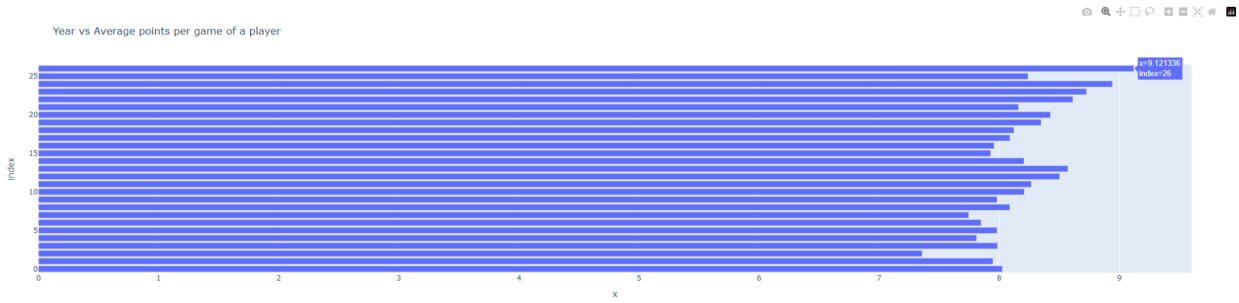


Figure 4:



To explain these figures, we'll first explain the process I went through to get them. First, I asked three questions and set out to answer them. The three questions were:

1. What is the relation between height and the rebounding ability of an NBA player?
2. What is the relation between weight and the rebounding ability of an NBA player?
3. What is the relation between how much the average player scores and year?

To set out to answer the first two questions I decided to create scatterplots of the data and make a linear regression fit. You can see this in figure 1 and figure 2. Both figures show a scatterplot of the data with a linear regression line drawn through the plot. On the y axis for both we see how many average rebounds a player gets per game. On the x axis we see height for figure 1, and weight for figure 2. After creating those two figures, I then used the metrics mean absolute error, sum of squared errors, mean squared error, root mean squared error, and root mean squared logarithmic error to evaluate how well the linear regression does based on a selection of five random players from the dataset. I also found the correlation for the two questions, which

ended up being less than I initially thought it would. The correlation between height and rebounding only ends up being 0.39848, while the correlation for weight ends up being 0.406723, which is only slightly higher. The closer a correlation is to 1, the better it is, with 1 being a perfect correlation. Since both of these correlations are around .4 though, we can see that there is only a very small correlation. This tells us that rebounding in the NBA only partially has to do with height or weight, and instead is likely more determined by player skill. The results of our metrics only help to further back up this idea, since our mean squared error came out to 0.6353236 for height, which isn't close to 0, which would be ideal because 0 would mean there is no error from our linear regression which means it could be used to perfectly estimate rebounds based on height.

For the last two figures, I set out to answer the third question: What is the relation between how much the average player scores and year? This question is spurred from the common idea that the NBA had changed a lot over the years, especially since it has gotten more scoring focused. Figure 3 is a bar plot that shows the average points per game of any given player vs year. This means we can see if players are scoring more or less since the average amount a player would be scoring would go up. Unfortunately though, our data shows that the scoring of the NBA hasn't changed drastically in this specific metric over the years. There are small changes here and there, but it doesn't move too far. Interestingly though, this still shows a large change in how many points are scored total in any given game over the years. Since the amount does go up over time a bit, every little bit that each average player scores more, the more any given team scores overall. This leads to show that while individual scoring hasn't gone up much over the years, scoring has gone up as a whole due to small point totals going up for each player on average. Figure 4 shows the same thing as figure 3, but it is interactive and doesn't quite look as nice due to me being less familiar with pyplot. In order to get the interactive plot to display the way I wanted it to, I was unable to properly give its axes titles. If I would have been able to, the y axis would have been 'Year' and would've properly

shown each year's name to the left of the bars. The x axis would've been 'Average points per game of a player', although the numbers on the x axis are as I'd want them. Looking at the figure we can see that the amount of points per game any given player is scoring reached its peak in 2022 (the most recent year in the data) at an average points per game of '9.12'. The interactivity of the graph can also be used to see how far the points per game scoring has come, since we can see that the first year of the graph has an average points per game for any given player of about '8'.

Conclusions/Where I Would Go in the Future

The main two things I got from my data were as such: rebounding in the NBA is more skill based than I thought, since player height and player weight have less of an effect on it than I first thought. This is pretty cool to know, since it gives me a greater respect for the players who are truly great at rebounding. The second thing is that points per game per player has gone up over time, but only a little bit. This results in teams scoring more per game which has led to the high scoring NBA that we have today, but it isn't as dramatic as I would've thought it would've been. I also learned that it is really cool to do analysis of data like this, since it can lead to conclusions that we would've never imagined before and it can teach us a lot about things we may have already assumed to be true.

As far as where I would go from here goes, I would look at the data and try to find some more good correlations to go into. I would then probably do something like I did with the first two figures where I create a regression model and then test to see how accurate it is. I would love to do this with some of the other stats in my dataset. I used a function within my project to look at the correlation of all number variables in the dataset. From that function I found two interesting pairs that I would look into if I were to do this again. I would look into the correlation between 'usg_pct' and 'pts', and I would look into the correlation between 'pts' and 'gp.' The correlation

between usage percentage and points is one that I think makes a lot of sense but would still be interesting to look into. It makes sense that a player who has the ball more (a higher usage percentage) would also score more. As far as the second one goes, it shows that there is a correlation between the average points per game a player scores and the amount of games they play in a season. I would guess that this correlation exists due to the fact that a player who plays more games is less likely to be injured, and players who are injured less are probably better at scoring. Thus, to answer that hypothesis, we could look into the correlation between those two variables. I am sure the information to be found would be interesting.