Neil Nguyen
Coen 242
Tao MW 5:10 PM - 7:00 PM
June 7, 2021

NBA Stats Report

Background:

The NBA has been around for 75 years now but only within the last 50 years (1979) did it start tracking analytics. What are these analytics? How many points, rebounds, assists, etc. made by players. These were really just raw numbers to embellish players and maybe attribute some numerical value to them. At its best it was used to determine whether a player was a good scorer, playmaker, defender, etc. It wasn't until 2009 that the NBA started tracking much more advanced statistics. It began using an advanced video system that tracked every player's movements 25 times a second. Now with this information we're provided with elaborate shot charts, full team play analysis, close ups of player's strongest or weakest moves, etc.

The evolution of data took the NBA by storm. Things like this heavily influenced how players, teams, and management approached the game providing much more study material to consume and improve from. One prime example is just a few years ago, general manager Daryl Morey of the Houston Rockets (at the time) broke NBA boundaries by surrounding the Rocket's franchise player (James Harden) with other players handpicked using advanced statistical analysis coupled with traditional qualitative scouting. This turned out quite successful and allowed the Rockets to remain contenders for some years.

Motivation:

The ESPN Youtube account regularly puts out videos with deep analysis of players and their shot charts, favorite moves, and more. I'm fascinated by specifically the shot charts and how this data is analyzed. Although they use the shot charts to demonstrate much more complex things, I myself sought to just prove something very simple using their advanced statistics: ***The closer you are to the basket, the more likely you are to score. Likewise, the farther you are from the basket, the less likely you are to score.*** In a basketball vacuum, like you at the gym shooting by yourself, this should objectively hold true. However, in the NBA and even in pick-up games with your friends there are countless other factors that come into play that will affect your shot percentage. I wanted to see if this would be visible in the data of the NBA's 2020-2021 regular season.

Experiment:

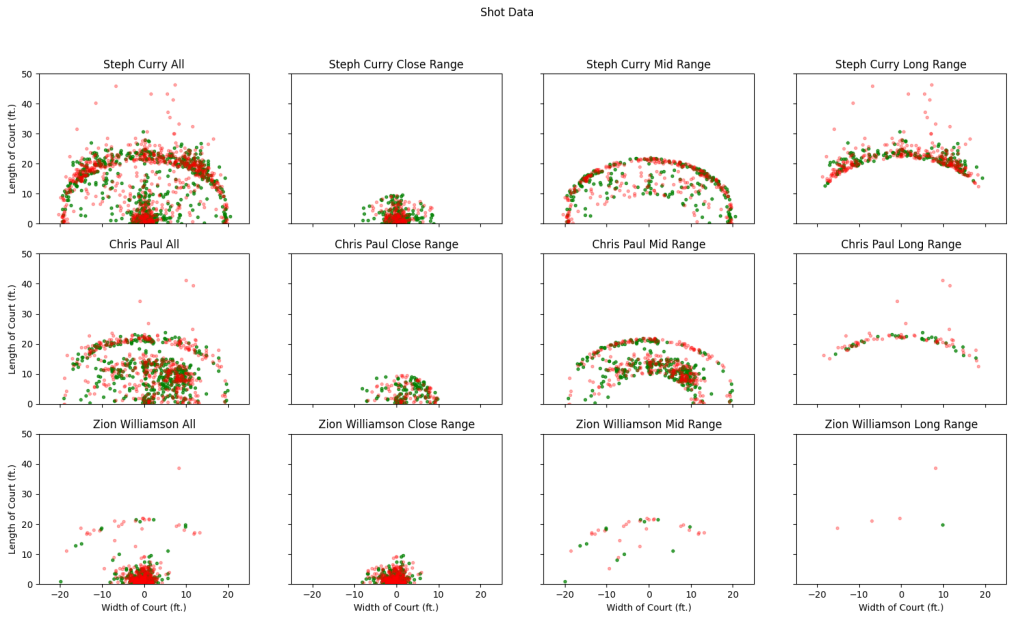
In order to prove my hypothesis I chose three players that were considered to be the best at their respective ranges this year: Steph Curry from long, Chris Paul from mid, and Zion Williamson from close. These ranges I chose as 22ft.+, 10ft.-22ft., and 0ft.-10ft., respectively. Ideally since these three are the best at these ranges they should be more successful at these ranges than others, right? That would, of course, go against my hypothesis and I aim to prove otherwise. Even the players at the best at a specific range might have an easier time scoring the closer they get to the basket!

Data Collection:

In order to collect the data I had to scrape from nbastats.com. I used a python file I labeled “stats_to_txt.py” to convert the stats I needed for my experiment into a text file that could easily be read as a list in Python or PySpark. How this was accomplished was by first taking all the shot data from a specific player’s page on the website (by ID#) and then converting it to a dataframe. After converting to dataframe I then removed each feature I deemed unnecessary to my particular experiment, things like opposing team, shot ID#, type of shot, etc. All I wanted was the x and y coordinate of where the player shot from, keep in mind that the basket is at (0,0). The x coordinates represent the right of the basket if positive or the left of the basket if negative. The y coordinates represent how far away you were from the baseline of the court. What I ended up with was a feature vector that had only 3 variables: x coordinate, y coordinate, and hit/miss flag (0 if miss, 1 if hit). Then I wrote this into an easy to read text file that Python or PySpark could convert to a list.

Graphing the Data:

I wrote “txt_to_graph.py” using pyplot that displays the feature vectors defined previously as graphs. Additionally I split up each range of each player into their own graphs. As you can probably tell it’s difficult to draw conclusions from the graph alone, which is why we must use PySpark/Rdd’s and Linear Regression in this experiment to prove my hypothesis. (Reds are misses, greens are makes)



Experimental Question 1: Do the players that are the best at each range (long, mid, close) have higher success rate at their respective range?

Using PySpark I wrote 5 functions to run my experiment using rdds:

1. txt_toRdd(sc, file): converts a player’s shot data text file to Rdd
2. rdd_toDistance(rdd): returns an rdd that maps the (x, y, hit/miss flag) rdd to (distance from basket, hit/miss flag) rdd using the distance formula. The distance from the basket represents the key while the hit/miss flag represents the value.
3. rdd_toRange(sc, lower, upper, rdd): returns an rdd that only contains shots within the lower to upper range provided.
4. rdd_toShootingPercentage(rdd): calculates the success percentage of all the shots in the provided rdd.
5. txt_toShootingPercentage(sc, lower, upper, file): bundles steps 1-4 all into one convenient function for data collection.

With these functions I collected each player's success rate at each range and showed that despite being the best at their respective ranges, it was still more likely they scored the closer they got to the basket. Here is the data below:

Player	Range	Count	Percentage
Steph Curry	All	1365	48.21%
Steph Curry	Long	457	41.58%
Steph Curry	Mid	490	43.88%
Steph Curry	Close	418	60.53%
Chris Paul	All	879	49.94%
Chris Paul	Long	93	32.36%
Chris Paul	Mid	577	50.09%
Chris Paul	Close	209	57.42%
Zion Williamson	All	1037	61.13%
Zion Williamson	Long	5	20%
Zion Williamson	Mid	37	32.43%
Zion Williamson	Close	995	62.41%

If you look at the percentages per player, the closer the range, the higher the percentage. However, some other things to note were that some of Chris Paul and Zion Williamson's volume of shots at other ranges were abysmal, not necessarily enough to draw concrete conclusions from.

Experimental Question 2: Using Linear Regression, does each player's success rate at incrementing ranges decrease?

I inserted every shot from each player into their own linear regression model from sklearn to answer this question. The dependent variables were the distances each shot was made from and the independent variables were whether or not the shot hit (0 for miss, 1 for make). Here are the important variables derived from the linear regression model for each player:

Player	Coefficient of Determination	Y-Intercept	Slope
Steph Curry	0.03319671697200777	0.6341533029088864	-0.00975059
Chris Paul	0.02513609802354977	0.6612750168414618	-0.01192462
Zion Williamson	0.053937840275777194	0.6777857364103181	-0.0280276

As you can see in the above table, the slope is negative for each player meaning that the linear regression model also shows that the farther away each player gets from the basket the lower their chances of scoring are. The Y-Intercepts the peak scoring percentages each player has at 0 feet. Further below I provided the predicted scoring percentage of each player at incrementing distances of 5. This data itself shows a decreasing pattern, also proving my hypothesis for these three players.

	0ft.	5ft.	10ft.	15ft.	20ft.	25ft.	30ft.	35ft.	40ft.	45ft.	50ft.
Steph Curry	63.42%	58.54%	53.66%	48.79%	43.91%	39.04%	34.16%	29.29%	24.41%	19.54%	14.66%
Chris Paul	66.13%	60.17%	54.20%	48.24%	42.28%	36.32%	30.35%	24.39%	18.43%	12.47%	6.50%
Zion Williamson	67.78%	53.76%	39.75%	25.74%	11.72%	-2.20% ~0%	-16.3% ~0%	-30.32 % ~0%	-44.33 % ~0%	-58.35 % ~0%	-72.36 % ~0%

The slopes show a steady decline in the percentages, with Zion's going much lower and even hitting the negatives (which were just rounded to 0%).

Experimental Question 3: Using logistic regression, can we predict whether a player will make a shot at incrementing ranges?

In order to build an accurate logistic regression model I first had to take the original text files I made prior and convert those feature vectors to pairs of (distance, make/miss). Then I normalized these distances by dividing them by 100, to keep them within the interval [0,1] I used 100 because that's the highest distance a player could shoot from but it's worth noting that no player even attempted shots close to that. Then to improve the accuracy I feature-lifted the input (distance) from [x] to [x^4,x^3,x^2,x^1,x^0]. Now with these new feature vectors I used train_test_split to build a logistic regression model for each player and calculate their accuracies, you can see these below:

Player	Training Data Size	Testing Data Size	Accuracy
Steph Curry	1092	273	53.48%
Chris Paul	703	176	52.84%
Zion Williamson	829	208	65.38%

Now using these newly made regression models, we can attempt to predict whether or not a player will make a shot at each incrementing range, here's the table below:

Player	0ft.	5ft.	10ft.	15ft.	20ft.	25ft.	30ft.	35ft.	40ft.	45ft.	50ft.
Steph Curry	1	1	0	0	0	0	0	0	0	0	0
Chris Paul	1	1	0	1	0	0	0	0	0	0	0
Zion Williamson	1	0	0	1	0	0	0	0	0	0	1

Strangely enough, the logistic model predicts Steph not making any shots farther than 5 ft. Maybe this is attributed to his extremely high volume scoring? And for some bizarre reason Zion can make the 50ft. shot and maybe it's because he has such low volume from far range? Besides the outlier of Zion's 50ft. shot generally the logistic regression shows that it's less likely to score from farther from the basket.

Conclusion:

In the end, all three of my experimental questions worked to prove my hypothesis. The PySpark Rdd's showed that despite being the best at their specific ranges, each player shot better from closer to the basket. The linear regression model showed a downward trend in shooting percentage the farther each player shot from. The logistic regression model predicted less shot makes the farther a player got from the basket with some strange outliers which can be dismissed due to the models' ~50% accuracy.

In the future if I were to revisit this project I would try using every player's shot data instead of three best at each range. Additionally, I should consider other factors besides shot distance like opposing defense, clutch time, team match-ups, etc. I also found out that the high-volume variety scorers seemed to provide more useful data (like Steph).

References:

- Kopf, D. (2017, October 18). *Data analytics have revolutionized the NBA*. Quartz. https://qz.com/1104922/data-analytics-have-revolutionized-the-nba/
- *Official*. (n.d.). NBA Stats. Retrieved June 8, 2021, from https://www.nba.com/stats/
- Teo, V. A. P. B. D. (2020, February 3). *NBA Shot Charts Part 1: Getting the Data (Python)*. DataVizardry. https://datavizardry.com/2020/01/28/nba-shot-charts-part-1/