Vatsav Sethupathi and Jabin Tome

DATA 363 Final Project Report

Is an NBA player's three-point shooting ability a factor of their height?

Introduction

The National Basketball Association, or 'NBA' is an American professional basketball league. It was founded in New York City on June 6, 1946 as the Basketball Association of America. Currently, it is one of the four major professional sports leagues in the United States and Canada and it is composed of 30 teams, 29 of which are in the United States and one of which is based in Canada. From its conception, it has been a haven of professional basketball and hosts the world's greatest talents in the sport.

In the first 30 years after its conception, the NBA had only a single category of field goals, namely the 'two-point shot'. This constituted shooting the basketball from any part of the court into the opposing team's hoop. Despite this limitation, many all-time greats boast several records concerning the total number of points scored in a game, including Wilt Chamberlain, who still holds the record of the most points scored in a game, with an individual score of 100 points in a match. Now as a result of the surging popularity of the league in 1979, the NBA decided to introduce a new class of field goals titled the 'three-point shot'.

A three-point shot (or three-pointer) is a shot that is made from beyond the 'three-point line', a designated arc with a radius of approximately 7.2 meters. In contrast to the two-point shot, it stays true to its name and awards the team three points for a successful attempt. According to the George Mikan, the commissioner of the league at the time, the three-pointer "would give the smaller player a chance to score and open up the defense to make the game more enjoyable for the fans". Although this introduction into the NBA was initially seen as a gimmick to improve the viewing numbers of the

league's matches, the three-point shot stayed true to Mikan's predictions and served as a crucial weapon in the hands of players who were shorter than their peers.

In this project, we will be exploring the relation between the height of a player and their three-point shooting abilities. In other words, we will try to find a correlation between a player's height and their three-point scoring and test the claim that shorter players are better three-point shooters. In general, we expect to find that there is a negative correlation between the height of a player and their three-point shooting numbers, that is, the taller a player is, the worse their shooting numbers will be. This is assumed to be true since taller players typically tend to play close to the hoop to outmuscle the defense while shorter and lighter players take advantage of their speed and shoot far away from the basket. Although this fact is an accepted as an unspoken rule, it has never been proven. Therefore, we hope that our study contributes to the acceptance of this observed trend.

Methods Used

In our attempt to prove this claim, we will employ two methods to visualize and analyze our data, namely bar graphs and a linear regression model. The variables we will primarily focus on in each of these methods are the height of a player and their average career three-point shooting percentage. We decided that a player's shooting percentage is a better reflection of his shooting ability than the average number of three-pointers he makes in a match since the time spent by that player on the court doesn't play a role in the percentage. This contrasts with the number of three-pointers that a player makes in a match since this number increases with the amount of time they stay on the court. There are many players who are substituted in to increase the team's score by simply shooting a couple of three's and it would not be accurate if we gauge their skills based on the sheer volume of shots they make compared to a player who has the opportunity to play throughout the entire game.

For our first visualization, we will create a bar graph that maps a certain range of heights to the average shooting percentage of the players that fall into that range. We are using 8 unique 5cm height intervals starting from players who are less than 180cm to players taller than 210cm. To obtain this data in R, we managed to create a for loop that iterates through our data and stores the total sum of percentages of people that fall within a certain height range. After this, we divide each of these sums by the number of players that fall within that range to get the average shooting percentage of that entire group. Once we have this data, we display it in the form of a bar graph with the x-axis denoting the height of the players and the y-axis denoting the shooting percentage of the player.

For our second visualization, we will plot our average data as a scatterplot with the x-axis denoting the height of the player and the y-axis denoting their shooting percentage. With this, we plan to create a linear regression model which maps a player's height to their predicted shooting percentage. To perform this, we simply used R's built in lm() command to generate a linear model of our data. Finally, we will plot our line of best fit to see how well we are able to capture the real-world numbers that we get and to check whether we have a negative correlation between the variables we are measuring.

Finally, once we obtained a result that followed our initial hypothesis, we performed an ANOVA test on the mean shooting percentages of each of our height categories to make sure that the values of the means that we get are significantly different to each other. This was followed by multiple pairwise comparisons to check which groups differed the most.

The data we are using is sourced from a GitHub repository of a user who scraped through the statistics posted on the official NBA website. The conclusions section contains data that we managed to scrape by ourselves from the official NBA website to analyze the trend of three point shooting throughout history. The link to the GitHub repository is given in the References section of the report.

Results

Shown below is the visualization we received with our bar graph:

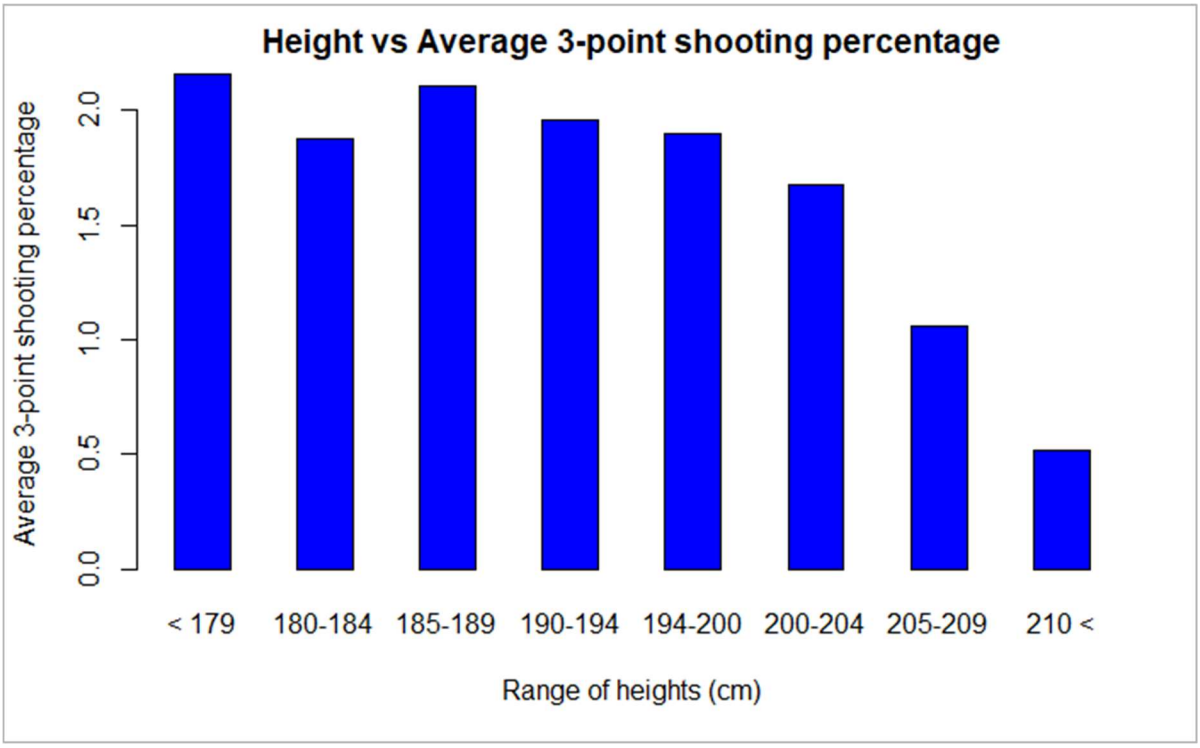

Figure 1: Bargraph of Height vs. 3-point shooting %age

As we can see, barring the 180-184 cm height range, we see a clear trend of decreasing bar heights as we move from left to right. The shooting percentages have a range of 1.6% with the highest shooting percentage recorded by players who are shorter than 180cm and the lowest percentage recorded by players who are taller than 210cm. Therefore, the general trend that is demonstrated by this graph is that on average, shorter players are much better when it comes to three-point shooting ability.

Next, a visualization of our linear model and its summary is shown below:

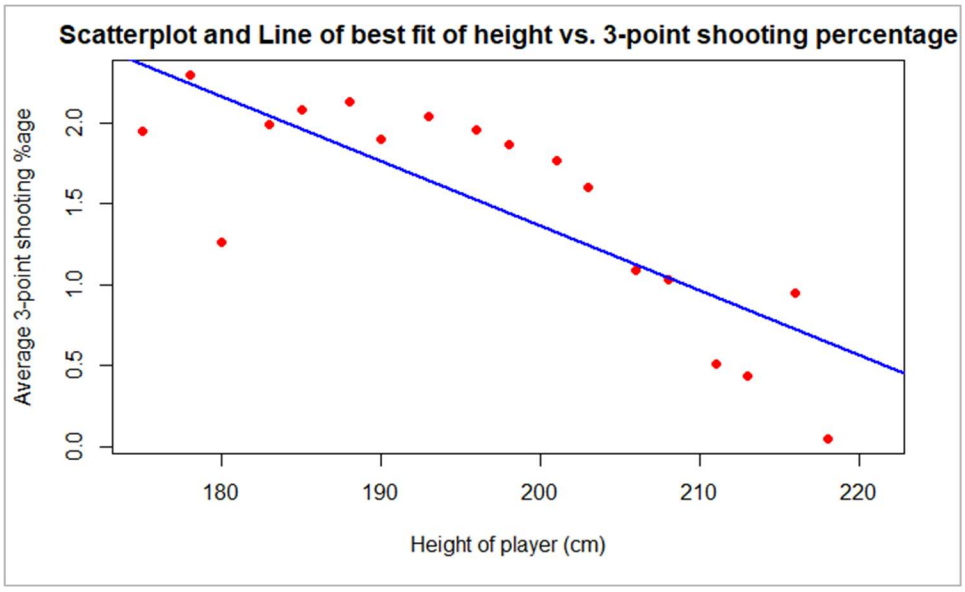

Figure 2: Scatterplot for height vs 3-point %age

When we observe our scatterplot, we see that the points roughly form a line. This would mean that there is a moderate correlation between the variables we have chosen. This observation can be confirmed by analyzing the summary of our linear model which will be analyzed further in the analysis section of the report. Our final visualization is a boxplot that contains the summary statistics of the shooting statistics for each of the height categories, which we will use for our One-way ANOVA test:

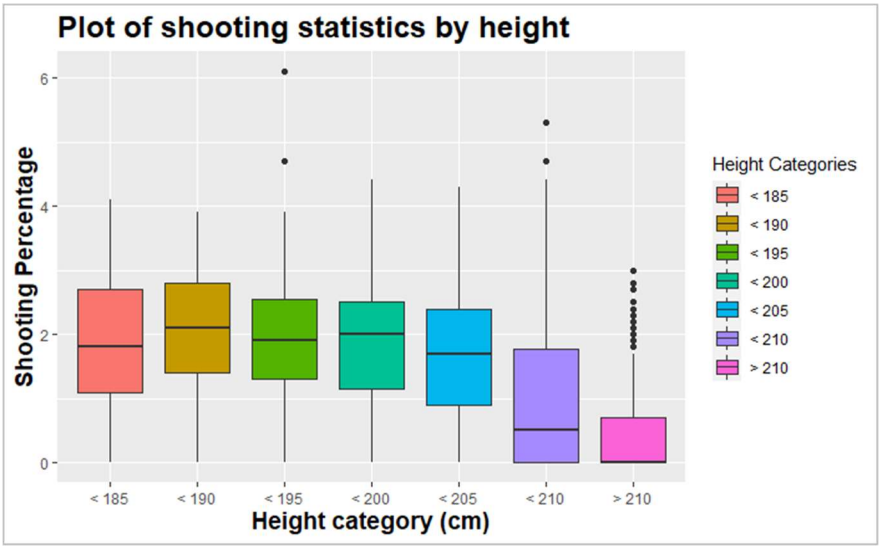

Figure 3: Boxplot for shooting percentage for each height category

Similar to our bar graph, we observe that the mean shooting statistics of the entire group follow a decreasing trend as the height category increases. One of the most interesting observations in the graph is the large number of outliers in the final category, namely the category for players above 210 cm tall. This can be explained by the mean being very close to zero, meaning a few exceptional players who shoot well despite being extremely tall end up becoming outliers. A few notable examples of such players are Dirk Nowitzki and Kristaps Porziņģis, both of whom are international players who play for the Dallas Mavericks.

Analysis

Given below are our summary statistics for our regression line:

```
Call:
lm(formula = newavg ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89713 -0.32317  0.08729  0.33734  0.43982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.337662   1.445122   6.462 7.86e-06 ***
x           -0.039855   0.007327  -5.439 5.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4099 on 16 degrees of freedom
  (29 observations deleted due to missingness)
Multiple R-squared:  0.649,     Adjusted R-squared:  0.6271
F-statistic: 29.58 on 1 and 16 DF,  p-value: 5.46e-05
```

Figure 4: A summary of our linear model

For our regression line, we see that our slope is -0.04. In the case of the data we are analyzing, this means that for an increase of 1cm in height, the average shooting percentage decreases by roughly 0.04%. This coincides with our original assumption that the shooting ability of a player decreases with their height. Moreover, we can be confident that the values that we received for our y-intercept and

slope are very accurate since the measures for Pr(>|t|), or the p-value for these estimates, in both cases are quite close to zero.

Finally, we see that our R^2 statistic has a value of 0.6271. We are considering the adjusted R^2 statistic since it accounts for the number of observations in our model and adjusts the correlation accordingly. We can interpret our value for the R^2 statistic as a measure of the linear relationship between our variables. This statistic takes a value between 0 and 1. Our model gave us a value of 0.6271, which signifies that we have a relatively strong correlation between our variables. This helps us conclude that there is a linear relationship between our variables, which in this case is negative.

Finally, to confirm that there is a significant difference between the mean shooting percentages of each of our groups, we performed an ANOVA test on our dataset. Our null hypothesis was that there is some difference between the means of any of the two groups. This would mean that the alternative hypothesis states that there is atleast one pair of groups that are significantly different from each other. The results we received for the test are as follows:

```
             Df Sum Sq Mean Sq F value Pr(>F)
group         6  208.1   34.68   32.07 <2e-16 ***
Residuals   653  706.2    1.08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: The summary statistics for ANOVA testing

Since we are comparing 7 unique groups, we see that we have 6 Degrees of Freedom. Now we notice that our F-statistic comes out to be 32.07. The larger the value of the F-statistic, the further to the right it is on the F-distribution. This would mean that the area under distribution for particularly high values of F would be very small, meaning it would be easier to reject the null hypothesis. In this case, our value of 32.07 is extremely high, meaning our area under the curve is close to zero. Another way we could confirm this is by checking our p-value. Since our p-value is almost zero, we can reject the null hypothesis at even the 1% significance level.

Finally, we wanted to find which of our groups are the most statistically different to one another. Therefore, we ended with multiple pairwise-comparisons using an inbuilt function in R known as the Tukey Honest Significant Differences test. Our results for the same are given below:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = value ~ group, data = boxstuff)

$group
                     diff        lwr         upr     p adj
< 190-< 185  0.17361538 -0.5749434  0.92217421 0.9933556
< 195-< 185  0.02561616 -0.6628387  0.71407105 0.9999998
< 200-< 185 -0.03498990 -0.7234448  0.65346499 0.9999990
< 205-< 185 -0.25888136 -0.9360694  0.41830665 0.9185651
< 210-< 185 -0.87572603 -1.5414648 -0.20998722 0.0021152
> 210-< 185 -1.41864463 -2.0943634 -0.74292585 0.0000000
< 195-< 190 -0.14799922 -0.6747697  0.37877129 0.9816919
< 200-< 190 -0.20860528 -0.7353758  0.31816523 0.9047701
< 205-< 190 -0.43249674 -0.9444544  0.07946091 0.1614268
< 210-< 190 -1.04934141 -1.5460558 -0.55262702 0.0000000
> 210-< 190 -1.59226001 -2.1022727 -1.08224737 0.0000000
< 200-< 195 -0.06060606 -0.4977754  0.37656323 0.9996329
< 205-< 195 -0.28449752 -0.7036997  0.13470465 0.4110124
< 210-< 195 -0.90134219 -1.3017857 -0.50089867 0.0000000
> 210-< 195 -1.44426079 -1.8610854 -1.02743622 0.0000000
< 205-< 200 -0.22389146 -0.6430936  0.19531071 0.6955246
< 210-< 200 -0.84073613 -1.2411796 -0.44029261 0.0000000
> 210-< 200 -1.38365473 -1.8004793 -0.96683016 0.0000000
< 210-< 205 -0.61684467 -0.9975919 -0.23609742 0.0000421
> 210-< 205 -1.15976327 -1.5577032 -0.76182335 0.0000000
> 210-< 210 -0.54291860 -0.9210465 -0.16479066 0.0004944
```

Figure 6: Results of TukeyHSD test

From this test, we see that most of groups' statistics resemble one another quite a lot. This can be inferred from the extremely high p-values for some of the pairs, for example the value of 0.9933 in the very first row. However, we received an extremely low p-value from our ANOVA test because of how different the shooting statistics are for the group with height over 210 cm. We see that for every other group compared with the final group, the p-value we get is very close to 0. Therefore, this contributes to the extremely low p-value in our ANOVA test since although most of our groups are distributed significantly similarly, the final group acts as an anomaly since it is completely distinctly distributed.

Conclusion

From all our testing, we can conclusively state that the three-point shooting ability of shorter players is significantly better than that of taller players. An interesting observation, however, is the lower height of the bar for players in the 180-184 cm height range. An explanation for this range could be the low number of players that fall within this height range itself. Therefore, even if a single player performs poorly in this category, they would bring the average of the entire group down. Another possible source of bias in our study is that the statistics we procured are the all-time career statistics of each of these players. Therefore, it is highly likely that our study could be affected by the fact that not a lot of players scored three-point shots in the 80s and 90s. However, we assume that this bias is balanced out by the fact that most of the players whose data we've collected are from the modern era of basketball, and that their individual contributions to the results we obtained are quite minimal.

Finally, we want to end with a visualization that demonstrates how the usage of the three-point shot has increased throughout history:

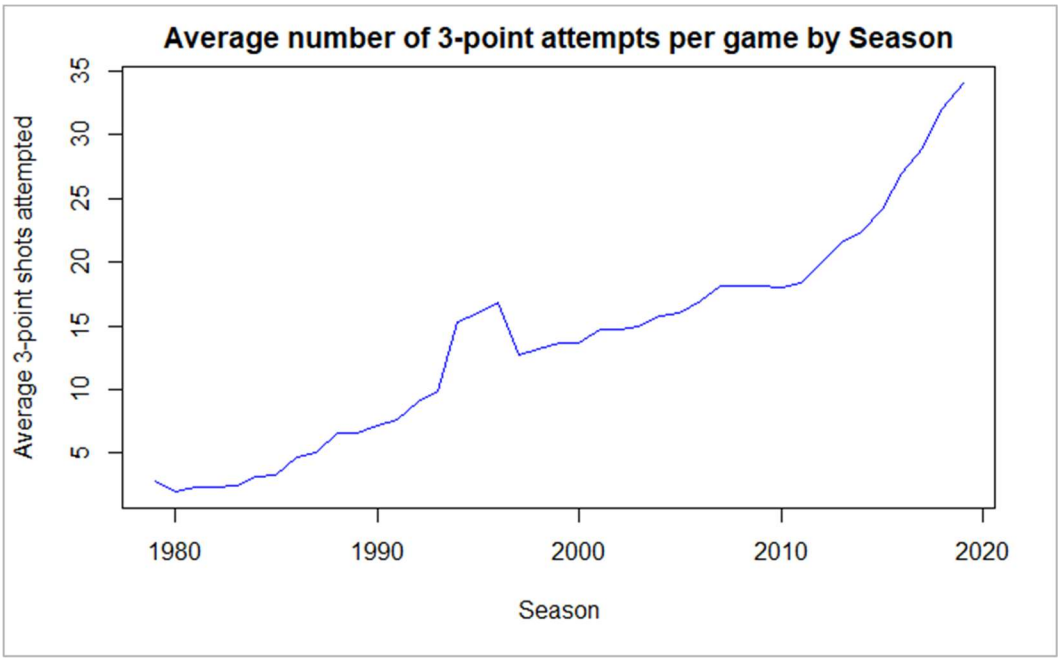

Figure 7: The average number of 3's attempted in a game per season

We see that since its inception, the popularity of the three-point shot has been increasing, now growing almost exponentially in the recent seasons. This exponential growth starts at around 2011, which happens to be the year that Steph Curry, a player who is 185 cm tall, had his breakout season. Soon after, teams started to notice the untapped potential that they had in their hands and started bringing in shorter players who could shoot the ball far away from the net. Therefore, although it may have started off as a gimmick to improve viewer ratings, the three-point shot has managed to cement its place in basketball and has added a new layer of depth, strategy and inclusiveness that contributes to its wide popularity and success.

References

Sources for data collection:

https://github.com/TWanish/NBAPlayerValue/tree/master/data

https://www.kaggle.com/drgilermo/nba-players-stats

https://www.basketball-reference.com/leagues/NBA_stats_per_game.html