An Analysis of NBA Scoring, 1980-2024

Watson Li, Brayden Butt, Frayan Foroughi CMPT 353, Fall 2024

With NBA player contracts ever growing and viewership falling short of other sports leagues such as the NFL, the NBA is faced with decisions to make. In an attempt to generate more revenue, the NBA has recently introduced the NBA cup which first appeared in the 2023-24 season. Adjustments have also been made to All-Star weekend to attempt an increase in viewership. One of these changes include the celebrity All-Star game, more notably, the addition of the 4pt line on the court during these games. Although this does not impact any real NBA games, it is perhaps a stepping stone towards a major change in the coming seasons. Could the addition of a 4pt line increase viewership and generate more revenue for the NBA? How would this change actually impact the games themselves? That is the question we aim to answer in this report: how would a 4pt line affect scoring in the NBA?

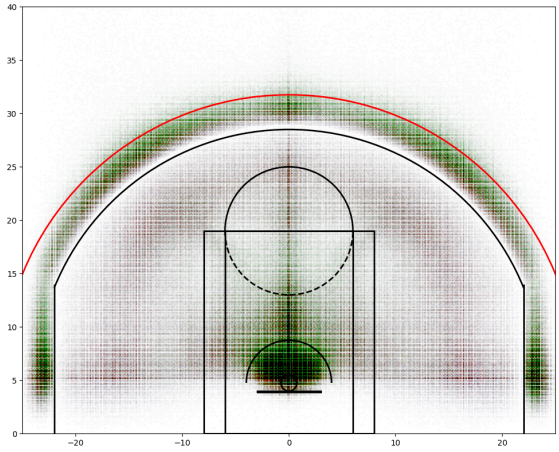
Although the first NBA season was in 1946, we chose to start our analysis in 1980, when the 3pt line was introduced. One data set included shot attempts from 1980-2003 while the other one was from 2004-2024 which included shot locations. Out of all the datasets, we chose to remove 1999, 2011, 2020, and 2021. The 1999 and 2011 seasons were removed because of lockouts which caused the season to be shortened and have a discrepancy in the data. This was because of a labour dispute between the NBA and the players. The 2020-2021 season was not included because of the COVID-19 pandemic which caused a decrease in games played and also led to some data being misrecorded.

For the 2004-2024 datasets we focused on the columns with GAME_ID, SEASON_1, TEAM_NAME, SHOT_MADE, SHOT_TYPE, LOC_X, LOC_Y, SHOT_DISTANCE, MINS_LEFT and SECS_LEFT. We created a for loop to consolidate the data from multiple csv files into a single pandas dataframe. Once everything was combined, we checked for missing values, typos, and data ranges of numeric variables. The dataset had no missing values or typos - however, we had to make an adjustment to TEAM_NAME variables since some of the team names were outdated, i.e. the New Jersey Nets had rebranded to the Brooklyn Nets. As for the data ranges, the values were mostly correct except for the LOC_Y values for the COVID seasons which were recorded inadequately and could not be used.

When descriptions of the data were brought up, a discrepancy was noted in the 2020-2022 datafiles. While the remainder of our 2004-2024 season data was fully usable after cleaning, the 2020-2022 data had LOC_X and LOC_Y values for the (X, Y) coordinates each shot was made from that differed from the rest. While the 2004-2019 data and the 2023-2024 data had X values ranging from -25 to 25 and Y values ranging from 0 to 94, the 2020-2022 datafiles had X values that were one-tenth (10%) of the other datafiles. This would have been easily correctable by scaling, but the major issue was the way the Y values were presented: these

ranged from 5.255 to 14.555, and a simple scaling method was unable to match them to the values in the rest of our seasonal data. As we were unable to reasonably scale the shot locations to match our other years, we left these years out of visualizations and calculations that required precise shot location data. However, we were still able to use the other values from these files after cleaning, namely, shot distances, makes, misses, types of shots, game IDs, and the like.

We began by creating a scatterplot of all shots made from 2004-2024, excepting the aforementioned years, and overlaying court lines atop the plot. When initially plotting the data as-is, we found that the observations were too dense to make out any detail, and we just got a solid blue plot. Matplotlib arguments were passed so that not only were our data points scaled down, but each would also be dropped to an alpha of 0.005, making individual points very close to being transparent, and simultaneously making it easier to identify dense clusters of data points. Even after this, however, we wanted more information, and regenerated the scatterplot, this time opting to color code by the season each shot was taken in, with older seasons in a redder hue and newer seasons in greener colors.



From this visualization, we were able to identify several major “zones” players tended to make shots from, with the heaviest concentration being the restricted area. Two other clusters were found for corner 3s, and a higher concentration was also observed in line with the hoop. Notably, almost no shots were observed immediately bordering the 3pt line; shots on either side tended to be separated by a clear distance, with the exception of the clusters in the corners, likely due to the tighter spacing between that part of the 3pt line and the court boundaries, players did not want to risk stepping out-of-bounds. However, the gap along the rest of the 3pt line can likely be attributed to players wanting to clearly demarcate their shots as 3pt attempts.

We also overlaid a red line representing our hypothetical 4pt shot line at a radius of 27 feet from the rim. From this line, we’re able to see that while the majority of shots lie well within the line, there is still a notable number of shots that may be classified as 4pt shots. Accordingly, our next step was to investigate whether or not shooting patterns have changed over time to increasingly favor higher point value shots. The easiest way to do this was to make another scatterplot, this time with a colormap corresponding to the season. However, matplotlib’s selection of native colormaps is limited, and our choice of the BRG colormap was not particularly visually appealing. While we were able to observe that in more recent years, shots have been being made from increasingly far distances and fewer shots have been made from

between the 3pt line and the key, we wanted a better visualization that could provide us with more detailed information.

As a result, we made another visualization, making use of the matplotlib library's ability to create gifs from a series of plots over time. Sections of the court were broken into five foot by five foot bins. Again, data from 2020-2022 was dropped from this visualization due to unusable data, so the years visualized were 2004-2019, and 2023-2024. Across all the years analyzed, we normalized the number of shots made from each bin per season from 0 to 1, with the highest bin getting a value of 1. A heatmap was then made for each individual year, with each bin for that year being assigned a shade of blue with its intensity corresponding to its normalized shot count value. When played back, this gif made it clear that as the years progressed, shots are being taken from increasingly further distances, as the areas around the 3pt line progressively darkened.

Once the data was visualized, we opted to simulate data based on the 2024 records to see how the 4pt line would have affected the scoring in the 2024 season. To complete the simulation, we needed a few key statistics that were not directly available to us with the datasets. Statistics that we mutated the data to acquire included 2pt FGA, 3pt FGA, 4pt FGA, 2pt FGM (field goals made), 3pt FGM, 4pt FGM, FTA (free throw attempts), and FTM (free throws made), all on a per-game basis. With this data, we also computed success probabilities by dividing the makes by the attempts. Now that we had the appropriate data, we were ready to complete an 82-game simulation. To do this, we used the binomial distribution to predict makes based on each scoring type's attempts and their success probabilities. Then by scaling each scoring type by their respective value (ie. 1, 2, 3, 4) we were able to obtain simulation data for an 82-game season. To come to an appropriate conclusion, we took an average of the 82-game sample and divided that by 2 in order to obtain the average points-per-game (PPG) of each team. The simulation provided an average points-per-game of 132.152439. However, scores in basketball are whole numbers, and therefore we will round down and say that the average points per game for each team based on the 2024 season would have been 132. The actual PPG per team in the 2023-24 season was 114.8, which we will round to 115. This means that if shots taken beyond our proposed 4pt line would have been counted, we would have seen just under a 15% increase in PPG for each team.

Additionally, we decided to just look at some simple statistics to see just how viable 3pt shots are in comparison to 3pt shots. Some quick calculations show that while from 2004-2024, 3pt shots were successfully made on 34.95%, or just over a third of attempts, shots from beyond our hypothetical 4pt line were only successful 23.83% of the time, or less than a quarter of 4pt attempts. The number of shots made from beyond the 4pt line are also notably less than the full number of 3pt shot attempts: successful shots that could be considered 4pt attempts only make up 4.6% of successful 3pt shots, and a measly 1.61% of all 3pt shot attempts.

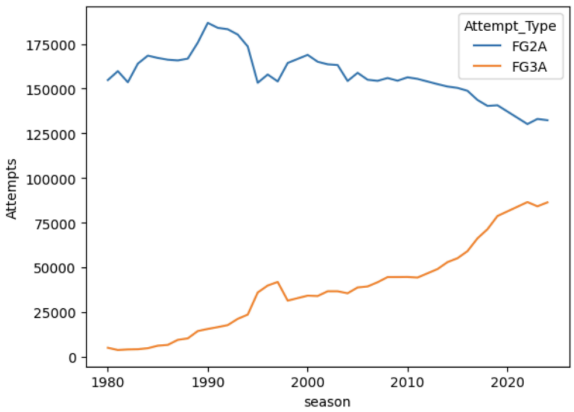
We then decided to take a closer look at just the data recorded for the 2024 NBA season to see exactly how much of a difference the 4pt line would have made in the NBA if it had been introduced in the 2024 season. Any 3pt attempt made from behind the 4pt line was thus reclassified; we kept the old values in a new column in order to calculate a score breakdown based on current scoring rules, as well. Aggregating the values, we managed to get scores for each game played by each team for the 2024 season according to both the proposed and the existing scoring mechanisms, but found that the scores themselves did not change much: for many games, there was no difference in score whatsoever, and the largest improvement in score was only 8 points, which was recorded in only 3 games.

As individual games didn't show much of a difference, we instead turned to examining individual teams. Aggregating the total number of 4pt shots made per team during the 2024 season, we found that the Golden State Warriors would have seen the greatest improvement in score, as a whopping 147 of the shots they made would have added an extra point. They were followed closely by the Milwaukee Bucks with 140 4pt shots made, but after this point, the number began to trail off; by the sixth most-improved team, the Philadelphia 76ers, the total improvement in score for the season was already below 100 points.

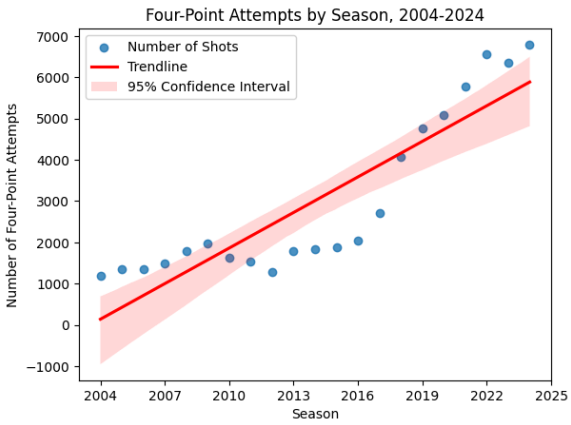
We also trained a K-nearest neighbours binary classification model to try to predict whether or not a 4pt shot would be successful using four predictors: the season a shot was attempted, the X and Y coordinates each shot was attempted from, and a fourth, engineered predictor, the time remaining in a quarter in seconds. This last predictor was created by multiplying the minutes remaining value by sixty and summing it with the initial seconds left value. While in initial evaluation, some significance testing suggested that reasonable predictions could be made without the use of the X coordinate, as its significance was merely marginal, the model was found to perform better with a 6% lower misclassification rate when the parameter was kept for training. Overall, the model was found to be able to successfully predict the outcome of a shot from beyond the 4pt line with just over 74% accuracy - not the greatest in the world, but much better than if it had simply been guessing randomly. These results suggest that perhaps there might be some importance in where players tend to attempt their 4pt shots from - whether that might be because some positions are harder to shoot from or are better defended remains to be known. Whether a player shoots towards the start or end of a quarter also has some relevance, but with the information at hand, we can't discern whether that's due to energy levels, adrenaline levels, or something else altogether.

After completing analyses on the recent years, it was time to look at the data over a longer period of time. The first step that we took was to use a t-test to compare means of 3pt attempts for each half-decade with the following half-decade. For example, was the mean number of attempts between seasons 1980-84 the same as the mean from 1985-89? This analysis would provide us insight into how attempts in the league were impacted when the 3pt line was

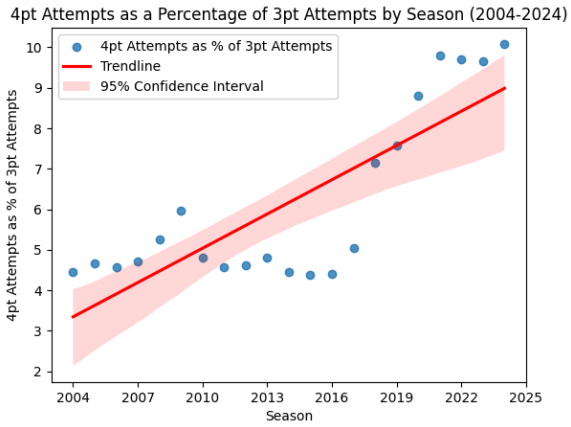
introduced. Although there is a possibility that adding a 4pt line would not provide similar trends, long-term impact could see similarities. With the t-tests, we were able to compute p-values to measure the significance for each "era". There was shown to be a significant difference in each era until the 1995-1999 seasons were compared to the 2000-2004 seasons. The following half-decade showed another increase, followed by the final comparison with no significance, namely 2005-2009 vs 2010-2014. The trend continued upwards showing a significant difference for the remaining years. If we chose to apply these trends to our analysis for 4pt lines, we would expect there to be a slow-start to 4pt attempts with a gradual increase. Spikes in attempts around 15 years after the implementation were often followed by plateaus, and in later seasons, the trend continued with a gradual increase. As previously mentioned, a number of seasons were removed from this analysis due to shortened seasons, namely, 1999, 2012, 2020, and 2021. If these years were included, this would result in a decrease in their respective eras, skewing the data and providing unrealistic results. If given more time, we could possibly look at adjusting the groups for the t-test to include shot attempts on a per-game basis, opposed to per season. Grouping the data this way would allow us to include shortened seasons. However, this analysis would not be perfect as the sample sizes would differ (ex. 2012: 66 games, regular: 82 games). Although grouping by game would allow us to include these values, it would be bad protocol to include seasons that were not gathered with equal sizes.



In addition to the 3pt attempts since 1980, we also wanted to know the breakdown between 4pt and 3pt attempts.Two regression models were fitted: the first was to assess whether or not we were seeing an increase in 4pt attempts over time. While we knew from the heatmap that shot distances were increasing, were players actually going further beyond the 3pt line before shooting? This first visualization was generated using a Seaborn regression plot, with a 95% prediction interval surrounding our trendline. To obtain the raw values for evaluation and hypothesis testing, a secondary model was fitted using the linear regression function from the scipy stats package. From the output, we found that the number of 4pt attempts was increasing by approximately 287 shots per season, with a standard error of about 34 shots and a 2 standard error confidence interval of [200, 346]. The p-value corresponding to this slope was 6e-08; well below our significance cutoff of 0.05. With this in



mind, we decided to make a prediction as to how many 4pt shot attempts would be made in 2030: while from 2004-2015, we pretty consistently saw only one or two thousand 4pt shot attempts per season, this number began increasing fairly consistently afterwards, and if our model is to be trusted, we forecast about 7600 shot attempts with a 95% prediction interval of roughly [5300, 9900] shot attempts. This is definitely a major increase over years prior! Now that we have a rough idea of how the raw shooting numbers would change, how would the number of 4pt shots change as a percentage of other shots over time? We decided to compare 4pt shots as a percentage of 3pt shots, rather than as a percentage of all other shots, as 4pt shots already only make up a small minority of 3pt shots as is, and it would be easier to evaluate changes in a larger percentage.



Once again, a Seaborn regression plot was used to create a visualization with a 95% confidence interval surrounding our trendline. What is interesting to note here is that the individual data points seem to follow a visual pattern very similar to that seen in our previous plot, suggesting a correlation between the number of 4pt shot attempts as well as the proportion of 4pt shot attempts relative to 3pt shot attempts. This could further suggest that players are increasingly opting to take shots from a further distance, and that players are intentionally aiming for more of these higher-stakes shots as the years progress. Breaking down the data obtained from the scipy linear regression on the same data, this time we see that the number of 4pt shot attempts as a percentage of 3pt attempts are increasing year over year at approximately 0.28% per year, which doesn't sound like much, but is actually fairly notable. The 2 standard error confidence interval calculated for this slope is approximately [0.24, 0.33]. Plugging some numbers into this model for the year 2030, we expect a whopping 10.68% of 3pt attempts to be from beyond the 4pt line, with a 95% prediction interval of [7.51, 13.85]. This is essentially a doubling compared to the 4% or 5% we'd previously see from 2004 all the way through 2016, with the exceptions of 2008 and 2009 which saw rates of between 5% and 6%.

The greatest limitations we found were the quality and amount of data and the amount of time available. Had we had more comprehensive data, we would have liked to make more accurate predictions about future shooting trends, and would not have had to drop certain years from our data, such as the seasonal data from the COVID pandemic with bad X and Y coordinates. However, modelling this would have needed even more data than what we had access to here, as well as significantly more computation. Even assuming a constant number of shots, the amount that players are incentivized to shoot from beyond a 4pt line rather than simply from shooting from the 3pt line remains to be seen, and as such we don't have a good estimate of how much to increase the proportion of 4pt shots by should the 4pt line actually be implemented,

even though we know the proportion is growing as it currently stands. Another example of unknowns would be the condition of individual players - mental states and further individual health breakdowns would have meant we could have focused more on individual players and their shooting habits rather than teams and making estimates, and we could have utilized more concrete data rather than simply performing simulations.

In conclusion, following our evaluations, we determined that there is a steady increase in 4pt shot attempts over time, and if the NBA were to actually implement the 4pt line, we would see an increase of scoring in games. Scores would increase most for teams who have shooters who specialize in long-distance shots, such as the Golden State Warriors with Steph Curry, or the Milwaukee Bucks with Damian Lillard. In the short term, these teams would benefit the most from the change, however, as shown with our statistical tests, we would expect there to be a significant increase in scoring long-term. As we saw the major increases with 3pt attempts and the increase in proposed 4pt attempts, we expect the rest of the league to catch up to the top teams in time. Even in recent years where we saw centers and power forwards in previous eras not attempt any 3pt shots, it is now expected of them. We could expect the same trend with all players with the addition of a 4pt line, as it would become expected that a player has the ability to make a shot from 4pt-distance, as opposed to the luxury of simply being able to.

Project Experience Summaries:

Watson Li

NBA Shooting Progression Analysis November 2024

- Reviewed and revised code in alignment with Pythonic best practices, and eliminated bugs
- Generated a variety of visualizations including scatterplots, animated heatmaps, and regression plots to evaluate the evolution of shooting practices in the NBA over time
- Calculated shot breakdowns to determine viability of four-point shots
- Performed regressions on four-point shots versus all shots and three-point attempts by year to assess significance of change in shot style
- Predicted increase in number of four-point shot attempts by 2030
- Trained K-nearest neighbours machine learning model to predict whether a four-point shot is a make or a miss based on season, shot location, and time remaining in a given quarter of a game
- Communicated relevant procedures and findings pertaining to code written in report

Brayden Butt

NBA Shooting Progression Analysis November 2024

- Retrieved, cleaned, and mutated data from Kaggle using Python
- Revised Python code to remove errors and ensure accuracy across code chunks
- Generated plots showing the evolution of shot attempts per season in the NBA
- Performed t-tests to measure significance of 3pt attempts between eras
- Generated simulations from existing data to measure potential scoring changes with a proposed 4pt line
- Communicated the problem and analysis of the project such that it was readable to all audiences

Frayan Foroughi

NBA Shooting Progression Analysis November 2024

- Retrieved and cleaned data from Kaggle using Python to understand data better and produce statistical responses.
- Discussed how data was retrieved and cleaned using concrete explanations to give readers a great understanding of the topic.
- Edited final report using conventional writing methods to inform readers on the problem at hand and the solution behind it.

- Verified the python code by running it within Python making sure that the code ran properly to ensure a successful program.
- Provided insightful feedback to teammates with ideas to improve current plans that improved quality of work.