Application of Big Data and Data Analytics:

Impact of Data and Statistics on NBA Basketball

Jacob Mitchell

The University of Akron

Abstract

Data science and big data are taking the world by storm. Data is collected, analyzed, and used to make business decisions, solve problems, analyze patterns, and much more. To further research data collections, tidying, and analytics, my senior research project is going to revolve around big data, its impact on the world, and spend time analyzing data with an intricate focus on NBA basketball. Ultimately, I will be collecting and organizing data and finding trends and seeing if I can use the collected data to make intelligent predictions. In doing so, I will learn more about collecting data, analyzing trends, and just scrape what the world of big data is possible of helping us accomplish.

**Application of Big Data and Data Analytics: Impact of Data and Statistics on NBA Basketball**

In today’s technology centric society, the world is always buzzing around the next big advancement and new adaptations of current technologies that can help streamline our society into one that’s more technologically efficient. Throughout the 2010’s we’ve seen dozens of advancements in technology such as the smart phone, the internet of things, and even prototypes of self-driving cars; these advancements in technology and our everyday use of the internet is more apparent than ever. As a result, there is more data stored about every facet of our lives that is expanding exponentially as time moves forward. With all this data and a never-ending drive to keep advancing our technologies, the field of data science is more prominent than ever before. Big data can be defined in a multitude of ways depending on profession, however as elaborated by Cody Agnellutti in *Big Data: An exploration of Opportunities, Values, and Privacy Issues,* “most definitions reflect the growing technological ability to capture, aggregate, and process an ever-greater volume, velocity and variety of data…More precisely, big datasets are ‘large diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.” (Agnellutti, 2014) The field of data science, although examples of people using data and statistics to solve problem ranges back more than hundreds of years, is coined as a relatively new term and its exact meaning can contain a bit of ambiguity or be used in several different contexts. This is further elaborated by Angelluti stating “both the technology of big data and the industries that support it are constantly innovating and changing….Big data may be viewed as a property, as a public resource, or as an expression of induvial identity. Big data applications may be the driver of America’s economic future or a threat to cherished liberties.” (Agnellutti, 2014). All technology has privacy concerns, but for the purpose of this research project, I will be setting those concerns to the side and analyzing the impact of big data and data science as a field. More specifically, big data’s impact in the world of sports at large with a focus on NBA Basketball.

Sports are culturally impactful all across the world. Soccer being a worldwide phenomenon and a multitude of other sports ranging from hockey to American football are a staple of life no matter where you’re from. The inherent competitive nature of sports enables teams and managers to want to have a competitive advantage. With the world of data science, to some extent that’s possible. In sports, it’s generally common for spectators, coaches, and team management staff to express or announce who is believed to have an edge in a game. Sometimes based on overall team skill and very general statistic such as a teams record make it seem relatively obvious as to what the outcome of the match will be; the team who has a much stronger win loss ratio will more generally be predicted to beat a team who has a worse ratio. The problem with these types of generalizations however is that they are more frequently commented on for the purpose of entertainment and these assumptions are subjective in nature. With data being collected about all players, games, and seasons down to atomically small attributes such as even the location of and time of when a shot was taken, the data collected can be plotted and critically analyzed to help general managers and their teams see exactly which attributes can be claimed to win them games.

The National Basketball Association (NBA) has been collecting base stats since as far back as their inaugural 1946-47 season(NBA FAQ) and have started collecting advanced stats, which are defined by the NBA as “a more in depth way to look at a simple box score, and more accurately [evaluate] the skill and production of a player or team”(NBA FAQ), since the 1996-97 season. All of their statistical records have been digitally recorded and are available on their website as early as 10 minutes after the conclusion of a game. Additionally, all the data collected and used for this research will be mined and analyzed from this source.

In order to manipulate data and begin to analyze what kind of influence it can have on the structure of the game, first we must have some sort of data set to work with. In the honor of delving into data mining as a whole, a field I priorly have had no experience working in, I spent the beginning phases of this project understanding and learning how to scrape data off the web. I found a tutorial on learning how to mine data called *Hardwood Convergence* written by Dan Watson, a health care data analyst who has a history in data analytics and does basketball analytics in his free time. To do this, I used a python tool called Jupyter Lab. Firstly, I needed to get all of the packages and libraries that would be essential in this project. I installed python on my machine, installed Jupyter lab and created an environment to work in, and in order to scrape data from the web and perform some simple calculations on said data I installed a tool called BeautifulSoup4. After importing these packages, I was then able to load a URL of an NBA player’s, more specifically James Harden’s, game log into a variable that would contain the webpages content on it, which was composed of very messy HMTL web data. Data mining in terms of its history was not always quite this easy but using modern tools we can streamline the process of scraping this data without it being an outrageous task. BeautifulSoup, has an HTML parsing function and a “prettify” function, allowing us to transform that messy data we took from the webpage into clean cut HTML that increases its overall readability and structure. In looking at the structure of the HTML tables that contain the stats, all the values are contained in <td> tags and are appropriately labeled in said tags as to what is being held there. To turn this HTML in to useable data, a data frame can be constructed that holds all values of the information that is needed for analyzing. To load these values into the data frame, a simple loop can be preformed that fetches the text where the tag is equal to “td” and export those values into the constructed data frame. Having that data stored into the data frame and viewing its components, there is information that was stored that was still not formatted in a usable way. In terms of this scenario, empty columns were stored for games in which a player was inactive for a game, skewing pieces of the pulled data under inappropriate headers. To fix this, what is known as data cleaning needed to be performed. This cleaning can also be done with a simple loop stating if a field was left blank, take the value that is stored to the right and push it to the left. This process took out the blank spots in the data and pushed the statistics pulled under their appropriate headers. After pulling all this web data and then cleaning it into an accurate usable data frame, it can then be exported to a csv that can be used to begin performing calculations.

Regarding this project, I wanted to be sure that I familiarized myself with aspects of data science and data analytics and all of its components. For the sake of time and indulging more time in analytics, the only data that I scraped from the web was priorly outlined. Albeit not the end of my time working with structuring files and organizing the data I need to work with, the scraping of this data gave me a foundation on the structure of datasets and ultimately pushed my understanding of how to work with pulled data, which positively influenced the analytics that followed. Before going into analytics performed after this web scrape, it is needed to outline the other methods in which I gathered datasets to work from.

As priorly described, web scraping is an efficient method in retrieving data to work with. The user can maintain control of the structure of what is pulled and organize it into data to be used. Instead of scraping the information of thousands of players, teams, and games into a colossal amount of separate data sets, I found an API that handles all that work for the user. There is a widely used package, aptly named nba\_api, which is an API client for nba.com. This package allows for more accessible use to API Endpoints making all of that data that was priorly pulled a lot easier to access. This package’s extensive documentation shows the data it can pull, including a player dictionary, a team dictionary, box scores, fantasy scores, and a plethora of other stats ranging from larger scale season stats to an atomically small shot location for each player. This package allows for much more ease of access of beginning to analyze data and make readable charts to make educated decisions.

The analytics journey of this project began with the aforementioned web scraping of player stats. After loading scraped data into a CSV file, I created a new notebook in which to begin performing calculations in. The first of the plots that I made were made using matplotlib, a python library that allows for easily plotting graphs. In order to plot some simple line graphs, data was loaded into two x and y values being the season game number (the 2017 was the season of game logs that were scraped) and points respectively. With this line graph, you are easily able to see the variation in the player’s scoring on a game to game basis. Simply put, you can see hot and cold streaks as well as just how high or low of a number of points said player has put up. While this is interesting to see, it serves little in making informed decisions on playmaking. To make this data a little more useable, additional variables can be added to the graph to cross reference between the two given statistics and give the data plot more meaning. In this scenario, field goal attempts were plotted as a bar chart overtop of the already plotted points per game vs game number. With this added variable, you can see the variability in how many points were scored versus how many attempts were made to score. Most usually, the higher the number of field goal attempts, the higher the number of points would be scored that game, which would make sense. But what you can see through this plot is that it’s not always the case. James Harden is famously known not only for being a professional basketball player in the NBA, but for his ability to be able to offensively draw fouls. A plot like this can help show the variability of how things such as these fouls can positively impact his scoring despite not always having a high number of field goal attempts. A simple graph like this can show you just how your player is being effective even on a night where they may not be as offensively aggressive. While these numbers saved in a CSV file may not always look like much, they are able to tell the reader a story and make better informed decisions to construct new strategies and work on how they run plays or spread the floor.

Working with the nba\_api is a relatively different scenario. Where before, only a single player’s game log for a season was scraped, plotted, and analyzed, now thousands of game logs, player stats, and team stats can be plotted and analyzed with much more ease. In order to familiarize myself with this vastness of data, I began by taking a single player’s information and working with it. Obtaining this data was a multistep process. First, I made a team dictionary out of all the teams in the NBA using the teams module included in the package. Then, I exported only the data I was going to use, for this first case I used the Golden State Warriors, and put it into its own variable. I used simple python code to set the team ID number into its own variable as well. Once I had that, I then repeated the process for the player module, making a player dictionary and isolating Steph Curry’s player ID number. Having these IDs was essential because the API is constructed with all of its information being organized by these keys. With having these keys stored into variables, I then constructed data frames using another nba\_api feature called LeagueGameFinder. I used this method to load data about every action on court that took place where the player ID was equal to Steph Curry’s, allowing me to have all the data down to where the shot he took was for every game he’s played while playing for the Golden State Warriors. Having all of this data, I began constructing different plots and expanding what I had learned by adding some new variables and features to the mix. Similar to the first plot made about James Harden, I plotted a simple scatterplot this time of Steph Curry’s 3 Point Field Goal Attempts (FG3A) against his 3 Point Field Goals Made (FG3M). As to be expected, this showed a positive linear correlation, as 3-point attempts went up most frequently 3-point shots made would go up as well. To see more details about this information, we add more variables. Plotting using matplotlib, the size of a point can be used to show other statistics. So, in order to paint a bigger picture, the size of the plot point reflected the amount of points scored that game. The larger a plot point, the higher amount of points Curry scored that game. Additionally, the color of the plot points can be changed and used as variables to convey more information about the data points. In this case, the blue dots meant the game ended in a loss, whereas orange reflected a victory. In the figure constructed, you can see that the higher the 3-point attempts taken by Steph Curry, the higher a percentage the chance that the Golden State Warriors won the game.

To further explore Steph Curry’s production on the floor, I used the nba\_api to produce another plot that model’s a linear regression of some of his advanced stats. In this figure, I plotted his turnovers against his overall plus minus. Plus minus refers to how positive or negative the team did while Steph Curry was on the floor. Turnovers are a bad stat to have a lot of, referring to when a player loses the ball and loses their offensive opportunity to score. Surprisingly, although the linear regression was expectedly negative, the slope of the regression was not very dramatic, being close to 0, meaning that even in games where Steph Curry had a lot of turnovers, it didn’t usually have a dramatic impact on how well he performed during a game. This in part could be attributed to external sources. One factor that can play a role in this is that the more minutes a player has on the floor, the more likely they are to have a higher number of turnovers. This same logic can be attributed to their plus minus rating. The more minutes the player has on the floor, the amount of time they have to make a positive impact in points production or playmaking will also rise. Plotting these graphs show an interesting perspective on Steph Curry’s production during a match.

The last plots I ended up making were much more involved than the first ones and show a very different type of statistic. Like the last plot, I used nba\_api libraries and functions as well as the same player dictionary and game logs to create similar data frames. In this example, I used two very different type of players. I gathered all game information for former Philadelphia 76er Kyle Korver, one of the best 3 Point Shooters in NBA history, as well as Giannis Antetokounmpo, 2018-2019 seasons MVP and a contemporary power forward all-star who has spent his career playing for the Milwaukee Bucks. I deliberately chose to make these graphs for two very different types of players, so it is important to outline where they specialize. Although a versatile player, Giannis specializes in playmaking and scoring close to the rim. His impact on the floor generally comes from within the three-point arc and closer to the restricted area under the rim near the baseline. Conversely, Kyle Korver spends very little time near the rim and instead floats around the 3-point arc awaiting a well-timed pass as he can accurately shoot an open shot. After creating data frames for each of these players containing their game logs, I focused on two smaller data points collected by the NBA; they store the x and y values of every shot taken by the player while on the floor, and whether or not they made or missed. Using these values, I was able to create scatter plots of precisely where on the floor every shot these players have taken in their careers. I found on open source plot to draw out the dimensions of an actual NBA floor to scale over top of this scatter plot that I used to more accurately show where these shots were taken. These shot charts can be used to show where a player is more accurate on the floor. The higher concentration of the blue plot points portrays an area where the player more frequently scores shots, also known as their hot zone. In this chart, I used Seaborn to create the diagrams, another python tool for creating graphs and plots from data frames. For the sake of readability of the graph, I flipped the x and y axis of the plot points to accurately depict how the court looks and to emulate how shot charts produced by the NBA and ESPN more generally look like. In plotting these shot charts, I was able to get bar charts on the top of the x-axis and to the right of the y-axis that help to portray the frequency in which shots are taken in the location of the floor. Again, for readabilities sake, these bar charts help the reader see a better understanding of how many shots are in that cluster of data points. In viewing these shot charts, you can see a much higher concentration of Kyle Korver’s shot’s taken being around the three point arc with some mid-range jump shots sprinkled in. Alternatively, in Giannis Antetokounmpo’s shot charts there is a very large cluster at the rim, with a decent bit of midrange shots and some sprinkles around the 3-point line. Where a player like Kyle Korver differs from Giannis, is that there are points on the floor where you can see he’s more accurate. Although playing in high school as a left hander, Kyle Korver shoots in the NBA with his right hand and as a result, you see a higher concentration on the right side of the court, where he is arguably more effective. He also has hot spots at both corners of the court, where your shooting hand doesn’t impact the shooting accuracy as much. Using charts like these, a team’s management staff can understand where player’s need to be to increase their offensive production and understand where the most efficient spots their player’s can shoot from and write plays that enhance their strengths.

While these shot charts are effective at painting a picture of where your player is efficient on the floor, there are additional tools you can use in python to look at these charts in a different light. To further explore these shot charts, I was able to create a color map of the same plot points. Using a color map helps to visualize a broader hot zone for a player. Where the graph is redder is generally where a low percentage shot would be taken by a player. Conversely, as the color gets into a brighter yellow or white, that shot would have a higher percentage of going in as the shooter generally shoots more efficient from that zone. Using these tools, I was also able to make the same kind of heat map out of the same shot chart, but instead of being a color map, the areas are split into hexagonal areas in which the same logic goes. In this chart though, a higher percentage shot is shown by a darker circle, and the lighter the area portrays that less shots were taken or made in that particular location. Charts like these can be used to help coaches and players understand where they are more efficient on the floor and to write plays that can help put players into those efficient spots.

Originally, I wanted to try to use the data I collected and analyzed to try to predict outcomes of NBA games; I hoped to predict whether or not a player would have a good game, a team would win the game based on a matchup, where a player would take the majority of their shots, or certain plus-minus rating for players depending on who all is out on the floor at one time. However, since the outbreak of COVID-19 in the first quarter of the year, I was unable to make predictions as the NBA postponed the rest of the playing season amidst the pandemic. The worldwide shut down happened in the middle of my research and before I started performing analytics on datasets, so in place of predictions, I’m going to conclude by exploring the impact of data science’s uses in the world, business, and sports as well as a synopsis of this research project.

In today’s world, no matter the circumstance, data plays a significant role in how businesses operate and help the keep technologically advancing. While although I’m focusing on NBA basketball analytics, these same principles can be applied to data that’s kept by companies and other businesses alike. Similar to how an NBA team would wish to make informed decisions to help navigate themselves to maintain an efficient team that would drive fans, a championship, and ultimately revenue, companies wish to use big data and analytics to give themselves a competitive edge in the market. For example, Harshdeep Singh writes in his article *Using Analytics for Better Decision-Making* that “organizations spend considerable time analyzing consumer data and frontline monetization opportunities, [and] it is equally imperative to focus on improving productivity and performance. Data and analytics can play a huge role in reducing inefficiency and streamlining business operations….Using analytics to measure key performance metrics across areas such as operational excellence, product innovation, and workforce planning can produce calculated insights to solve complex business scenarios.” (Singh, 2018) With huge datasets contracted by companies and internet users alike, you can begin to see patterns in data plots that can ultimately lead to helping optimize your products and lead to entrepreneurial advancements and help maintain revenue in any business.

The NBA has used data extracting and analytics similar to how a business would in order to change the way teams play their games. Stephen Shea, the Associate Professor and Chair of Mathematics at Saint Anselm College in New Hampshire with a Ph.D. in mathematics from Wesleyan University, has explored how the use of data has transformed the game of basketball in his article *The 3-Point Revolution.* Shea outlines the use of stats influencing the rise in shots from the 3-point arc in the late 1990’s to early 2000’s stating “the three provides value in two ways. First, it’s an efficient shot. Over the last 20 years, NBA players have averaged 1.05 points per above-the-break 3 and 1.16 points per corner 3. In contrast, players have averaged just 0.79 points per 2-point attempt outside of the restricted area. In other words, 100 mid-range jumpers will provide 79 points on average, while 100 above-the-break 3s would provide 105.” (Shea, 2018) Using simple stats extracted from NBA season logs in this example prove that teams understand that taking more 3-point shots will ultimately lead to scoring more points, especially when working with efficient shooters. This can further be explored by a player’s shot selection as outlined by the priorly constructed shot charts. Shea explores farther how players can find accuracy on the floor by comparing players Draymond Green’s and Anthony Davis’s shot selection and field goals made percentages. He explains how in the 2018 season “Anthony Davis had a better field goal percentage than Draymond Green in both mid-range and behind the arc. If we considered each player’s efficiency on their jump shots (from both regions), we’d expect [Anthony Davis] to be more efficient….And we’d be wrong. Draymond averaged 0.87 points per jump shot while [Anthony Davis] only produced 0.82. Simply put, Draymond averaged 5 more points per 100 jump shots. [This is because] Draymond Green had a far better shot selection. Only 17% of Draymond’s jumpers were mid-range attempts. A whopping 72% of [Anthony Davis]’s jumpers came from mid-range. [Anthony Davis] is the better shooter, but Draymond took better shots.” (Shea) This simple analysis of statistic shows the outlined cold zones in mid-range jumpers explored in the shot charts I constructed and allows teams and players to make better decisions to be more offensively efficient, and mining and analyzing this has ultimately slowly but surely altered the way we play the game of basketball.

Overall, data mining, big data, and data analytics play a colossal role in how the whole world operates. Whether a business, a normal internet user, or an NBA basketball association, studying big datasets and finding trends can help lead to making better, more informed decisions that can lead to more revenue, a better business model, more efficient work, or a stronger basketball organization. Data analytics has transformed the game of basketball, and has allowed teams and players to achieve a competitive advantage at finding what strategies are more useful, such as shooting more three point shots, where players should take shots from, which players play better alongside other players, and much more. Investing my time into working with these data sets has given me a firm understanding on how to use easy to access tools to create meaningful data frames, plot data to paint a picture and tell a story, and make more informed predictions on how to tell the outcome of certain game stats. It has also significantly portrayed the importance of data mining and analytics to help create a sustainable and efficient world that makes well informed decisions.

Works Cited

Agnellutti, C. (2014). Big Data: An Exploration of Opportunities, Values, and Privacy Issues. Nova Science Publishers Incorporated.

NBA. FAQ. (n.d.). Retrieved from https://stats.nba.com/help/faq/

Watson, D. (2019, July 17). Welcome to Hardwood Convergence. Retrieved from https://medium.com/hardwood-convergence/welcome-fefe6180b8aa

Singh, H. (2018, December 1). Using Analytics for Better Decision-Making. Retrieved from https://towardsdatascience.com/using-analytics-for-better-decision-making-ce4f92c4a025

Shea, S. (2018). The 3-Point Revolution. (n.d.). Retrieved from https://shottracker.com/articles/the-3-point-revolution