ARE NBA PLAYERS WORTH THEIR CONTRACT?

TEAM MEMBER:

ETHAN TAUBMAN

NBA player contracts have been increasing rapidly, with players getting bigger contracts and extensions from the previous year. This has caused teams to be faced with tough decisions whether they should retain their players or pay them based on their performance which creates a lot of uncertainty about the player's future. This project helps analyze whether NBA players are overpaid or underpaid based on their on-court performance and which teams have gained or lost from these contracts. By merging salary data and advanced metrics I was able to create new performance metrics to help determine expected salary to see if they are truly worth their contracts. The end goal of this project is to determine contract efficiency across the league and understand team level effects of paying these players.

To get the NBA statistics from this season so far, I web scraped from NBAStuffer.com and to get the salary for each player I downloaded a csv from Basketball Reference. My original plan from my proposal was to web scrape from DataBallR to get the NBA statistics and Spotrac to get the salaries, but I changed it because I was getting forbidden URL which was a major challenge in trying to web scrape the data. I kept trying to find different sites like ESPN, NBA.com, etc. to web scrape the statistics of this season to continue my project but ran into the same issue of forbidden URL. Same came with trying to find a website for the salaries from only a certain amount of the website existing with all the salaries which made me ultimately download a csv to get the salary data. From the web scraping of NBAStuffer.com, I collected 499 data samples while on the salary table I collected 471 data samples. Once I combined the two data sets together, I got a total of 472 data samples collected.

The data cleaning process included several steps to ensure that the player's salary and statistics were merged and analyzed accurately. First, column names in the salary dataset were standardized by renaming Player to NAME and Tm to TEAM so the columns matched the player statistics dataset. All the salary columns that were unrelated to this current season were dropped, as the question focused on overpaid and underpaid players analysis at the start of the 2025-2026 season. After merging the two datasets together, player name inconsistencies were identified. Some of the names were different from having generation suffixes, international characters, or nicknames. The unicodedata packages was used to normalize the international characters and manual name corrections were used to apply the matching for the two datasets to be able to merge successfully. Then players without any statistics were assigned a zero because they have not played this season yet because of injury.

Once the two datasets were merged successfully, there were a couple of more cleaning steps needed to ensure proper and accurate data analysis. The next steps were to remove any duplicate or not necessary columns names like rank, team_x and current and change column names like Team_y to TEAM and 2025-26 to SALARY to create an easier column name. The salary column was cleaned by removing the dollar signs and commas, so the data was able to be used for data analysis. Rows with no data received a 0 which helped with giving inactive players their tag from having a 0 in positions gave them the tag of inactive to show they were injured. Duplicate data points were then dropped as well as giving salary amounts to players with statistics. Finally, the columns were reordered so the data was easier to read from having NAME, TEAM, POS, and SALARY to appear first to help with identifying players.

Once the two datasets were cleaned and merged, the final dataset was used to analyze NBA player contract efficiency for the start of the 2025-26 season. The first metric that was calculated was the performance metric. The performance metric was calculated from using a weight combination of basic and advanced basketball statistics. The formula included points per game, assists per game, rebounds per game, steal per game, blocks per game, turnovers per game, offensive rating, defensive rating, true shooting percentage, effective field goal percentage, game played and minutes played. The weights were assigned to reflect the importance of each statistic while also adjusting for players availability, durability, and performance by scaling games played and minute played. Once the performance metric was created, the efficiency metric was then created by dividing the performance score by their salary. This metric measures their performance as well as availability per dollar spent. This allowed for comparison between different contract sizes. Then summary statistics were generated to examine the skews and distribution across the league for each column. Z-scores were then calculated from the performance metric to see which players under performed or over performed based on the league average.

To calculate the expected salary based on the performance, a log-linear regression model was used. A log-linear model was used because NBA salaries are not linear but grew exponential with a lot of very high contracts with fewer smaller contracts. This would be able to capture the relationship between the players' performance and pay very well by capturing the environment of the NBA. Based on the log-linear regression model, the expected salary was then calculated based on their performance. Then finally the value metric was calculated based on the difference between a player's actual salary and their expected salary. The positive values show that a player is overpaid while a negative value shows a player is underpaid. All these calculated values were used to analyze to identify trends across the teams and players.

For the visualization, I thought it was best to filter out players that are truly out for the season as well as players that played very few games so far from injuries that occurred. The first plot that was generated was a bubble graph to show the relationship between salary and performance with the games played being the bubble size as well as color showing the contract value with red being overpaid players while green shows underpaid players. This visualization shows the wide league perspective on salary and performance. The visualization indicated that there were more players with overpaid contracts than underpaid contracts. Which shows that teams can allocate their money better with better contracts. The second visualization shown is a histogram showing the distribution of contracts that are overpaid and underpaid. This reinforces the bubble chart showing that more players are overpaid than underpaid.

The next chart generated was a bar chart showing the top 10 overpaid and underpaid players based on their generated value. The top 10 overpaid players showed players who played very few games while being on huge contracts with some surprises as well. While the underpaid chart showed more players which were underpaid based on their performance. This was likely due to players being on a contract year or still in their rookie year playing at a very efficient level. Then the next graph showed the position relationship with value. The medians fluctuated throughout the positions from going to values around 0 but still showing that players are generally overpaid, but with pure guards and forwards there were more outliers towards overpaying their players opposed to the mixed positions of G-F or F-C with less outliers.

The final graph shows the team value with all the players on the team. This showed mostly all the teams being overpaid with values all being over 0 expect one team which was the Denver Nuggets who had a negative value for their players. They were not that much lower than 0 but still under which makes their team more effective with their money.

The premise of my project was to identify if players were underpaid or overpaid based on the performance as well as identifying if teams are overpaying or underpaying their players to help with business decisions later. My hypothesis was being able to compare and analyze a mix between basic and advanced basketball statistics to the salaries of players, I can identify which players are overpaid or underpaid. Based on the metrics and visualizations, most of the bigger contracts of players are being overpaid, but most of these overpaid players have been injured and have not played a lot of games like Bradley Beal, Joel Embiid, Lebron James, Paul George, Anthony Davis, Kawhi Leonard, and Trae Young. These players also have been getting older in age and are likely degressing in performance as well. The two players that I was shocked on the list are Steph Curry and Giannis Antetokounmpo how have known and shown to be huge factors to their teams. The players that are underpaid are not surprising from two being Nikola Jokic and Shai Gilgeous-Alexander Walker who are both favorites to win MVP and have shown they are the best in league. While the rest of the underpaid players are players with increased roles with their teams and thriving in those roles. But on the team level it shows that teams are all overpaying their players based on pure value expect the Denver Nuggets who have been outperforming their salary which make sense with Nikola Jokic being such an underpaid player.

During this process of evaluating players, there were a couple of changes that were different from my proposal. The first one was where the data was obtained. Originally the data was going to be scraped from DataBallR and Spotrac but ran into the challenge of not only these two sites but the more sites that I tried to scrape from there was a forbidden URL which prevented me from scraping the data. Which led the data scraping process to be changed to NBAStuffer.com. The other challenge that occurred was that all the salary websites had forbidden URLs which led me to download a csv from the basketball reference website to obtain that data. This changed my whole project from my proposal. This changed how I originally thought of showing the goal of underpaid or overpaid players with different metrics like points per salary, salary per win share, and salary per game as well as going to a few seasons before. This changed to one season and calculating metrics like performance, efficiency, expected salary, and value which ended up working very well for my hypothesis. The visualization mostly stayed the same but with different axis to highlight the premise of the project and the data that was obtained. Overall, the premise of the project stayed the same, but the difference helped analyze the players better than what was written in the proposal.

If more time were presented or different data obtaining techniques were allowed, the project would have improved significantly. My value metric was very accurate on predicting underpaid players and most overpaid players, but there were some fixes that could have occurred to make players like Steph Curry and Giannis Antetokounmpo not appear on overpaid. The first was to obtain data from 10 to 15 seasons before getting the full picture of these players' development and how they help the team. The issue of getting multiple seasons is trying to answer the questions of where you start and how do you determine different eras with salary. The salary was different from 10 to 15 years ago, which makes the statistics that the same salary for different eras would make no sense. Depending on where you start you will not be able to capture the full picture of the players' career. This can create bias and unfair judgments to the player. The second would be if there was a statistic to show how much an impact a player makes on the court like how the gameplan is built around them or some players play styles that are not shown on the stat sheet or how many defenders a player pulls because Steph Curry himself pulls three defenders on him every possession which does not show up on the stat sheet. As well as some players have an impact beyond the stat sheet which is important to have but not shown. Finally, having wins, apron level, salary cap, and team performances would have been great to show how well a team is doing based on their contracts. This project showed the individual impact of a player but does not show the overall all the players have together on team performance. The most important thing in the NBA wins which is not shown on team value. If wins, cap availability, and cap usage were shown teams like the Los Angeles Lakers, New York Knicks, Cleveland Cavaliers, and many more would have been underpaid since they are winning games collectively which would show more of team impact.