Linus Buchholz

Sidharth Mallela (Sid)

Python for Astro: NBA Salary Project Report

Note: Data is from the 2022-2023 NBA Season

Why We did the project:

When comparing players and how well they play, and whether they deserve the salaries they receive, player salaries in the NBA are frequently discussed. Certain players, such as Klay Thompson, are paid far more than they deserve, while underrated players like Zubac are paid far less simply because they are not as good offensively. We chose to do this project to uncover the truths about how the NBA treats players through salaries. The point of our project is to use features of players like their DBPM (Defensive Box plus/minus), PPG (points per game), etc… and see the relationships they have between the players salaries.

How did we process and filter the data:

We essentially just filtered out the players with either low games played or low minutes played. It's to have a more accurate view on the average NBA player; a player who had a 10 day contract or a player who came up from the G League for one game would heavily offset the average data, since many of these small time contracts would lower the average salary. In addition, if there’s player injury at the start of the season, it would give an inaccurate wholesale account of a player throughout the entire season. We filtered the code through seaborn, going through a certain column and checking if the value was lower than the minimum amount of games played or minutes played.

Theories we used:

Linear\_model\_fit() was the main section of the code where we created a scatter plot for each feature of the player compared to the normalized salary of the player. We can see that the DBPM vs Normalized Salary graph was a lot more spread out compared to the BPM vs Normalized Salary graph which shows that defensive players have salaries that vary more compared to offensively oriented players.

How the function works:

The curve\_fit function in the scipy library fits the linear model to the data points, the optimize module is used. It accepts input parameters for the linear function, the data (x and y), and the initial guess (p0). The function begins with an initial guess (p0) based on the mean values of y and x for the parameters m and b. This initial guess is used to begin the optimization process. A for loop is used to create multiple scatter plots, using the features I put into a list like so [‘BPM’, ‘DBPM’, ‘WS’]. The function then outputs the optimized parameters, their errors, and the linear model's final equation. For the given data, the equation represents the best-fit line. Some of the equations include y = mx+b (slope formula), [mean(y)/mean(x), -2.5] (Initial guess of parameters formula).

Takeaway:

1. A player's salary can vary depending on factors such as age, offensive ability, and injury history.
2. Generally, if you score more points on the basketball court, you get paid more.
3. Sometimes, there’s outliers in salary such as with Klay Thompson, who doesn’t have the greatest PER, but still has the 10th highest paying salary of 10 million
4. Box Plus Minus was generally more accurate in determining someone's salary rather than Win Shares
5. The Clippers are the highest paying salary with 3/10 top highest paid on their team.
6. Generally, when a player plays well, he gets higher salary