Who was the most overpaid player in the NBA for the 2022-23 season?

I'm setting out to find out who the most overpaid player in the NBA is using a data driven approach.

Before we get into any data or statistical measures, we first need to think about value. What does it mean for a player to have value? I think many people would have the gut reaction to start pointing to a player's performance. Maybe they would reference points, assists, or steals. Maybe they would start to reference the more advanced stats that we have to evaluate players: Player Efficiency Rating (PER), Value Over Replacement Player (VORP), or any of the other numerous ways we quantify performance.

For this initial analysis, we will only be looking at regular season data. One could make the argument that the regular season is in some ways irrelevant to the ultimate goal of winning an NBA championship. That the relevance of the regular season pales in comparison to the postseason. They might be right, but for this initial analysis we will only be looking at regular season data for our start. It gives us the most players and the most games to reduce outlier statistics. For additional analysis, a direct examination of playoff data would be an interesting avenue to pursue.

However, it's also worthwhile to note that the NBA and its teams are businesses. At the end of the day, teams want to make money. This is usually correlated with a team's performance, and thus correlated with the performance of the players on the team. But not necessarily; we observe that a player's status, fame, and personal lives can play a role in the exposure and allure of a team. Lebron James is still an excellent NBA player, but is clearly a player entering the twilight years of his career. Yet, his allure is still such that many people want to go to games to see him play. His jersey sales are among the highest in the league. Perhaps another overlooked way James creates value for his team is that *he draws other players to the team* as well. Other world class players want to play with James because of his track record and status (think Anthony Davis). Obviously, these attributes elevate the value of James beyond just his on court performance.

Another thing worth taking a moment to discuss, is basketball's unique properties which make it an especially good subject for this kind of analysis. Among the major team sports, I can't think of a more position-less competition. In the US, we have football, baseball, hockey, and basketball. Football is almost the polar opposite of basketball in that the positions are exceedingly different, and the type of analysis we are going to conduct here would be near impossible. In baseball, most players hit, which is useful for comparison, but the pitcher is by far the most valuable player on the field and doing an action which no other players are going to record statistics for. Again, this makes for a difficult comparison of all players. Hockey might be the most similar (excepting the goalie), but still poses two linearly structured teams against each other, so defenders are much less likely to score than attackers. Basketball poses two more flat-structured teams against each other. What I mean by flat-structured is that every player plays both offense and defense. Obviously, there are still positions in the NBA, and some

players are stronger on defense while others are stronger on offense. However, the overall similarity of the players on the court allows us to compare them all as one in a way that just isn't possible in other sports. Basketball lends itself to a more blanketed analysis approach than any other sport.

However, for this analysis we will only be looking at on court performance as a way to assess a player's value. We have a healthy variety of statistics to evaluate performance which is a good start. In fact, we probably have so many statistics that it might be difficult to keep track of what everything means. I've included a table with all the stats and descriptions in the appendix, and I will offer brief explanations as they come up. Additionally, it may seem inherently obvious that the most overpaid player in the NBA would be a player with a large contract who gets injured early in the season and doesn't play a large majority of the games. Again, we want to focus just on the on-court performance of a player so we won't hold that against the player.

Now let's get into the data. For this analysis, I'll be using excel to evaluate the data. Let's start by getting our initial data and refining it. We want to only include players who have played a significant number of games for this previous NBA season. I'll define this as over half (there are 82 games in a season which makes half 41). We'll start by only looking at the 5 basic statistics recorded in the NBA on a per game basis. These are: points, assists, rebounds, blocks, and steals.

Straight away we run into some complications in just the initial data refinement. Some players play for multiple different teams in a season, which means we have to average their stats weighted on the number of games they played for their tenure with each team. Luckily, excel has a built in function to remove duplicates and the website we pulled data for creates a row of data which already sums and averages the necessary data.

Another complication of our data arises when we try to combine data for salaries and performance metrics. Unfortunately, basketball reference doesn't provide salary data on previous years, forcing us to source salary data from a different source (I chose to use ESPN). This provides a challenge because different databases will have some players that won't appear in both. Yet another challenge comes simply from the way that names are stored in the databases themselves. The NBA database on salary from ESPN has names stored with accent marks, while the NBA performance data from basketball reference doesn't have these. Unfortunately, this affects the many foreign players who play in the NBA such as perennial MVP candidate Nikola Jokic. Given an extended amount of time, it might be possible to circumvent this problem, but the return on investment here just doesn't make sense for this project.

All told, after applying these conditions and losing some unfortunate players due to the crossing database issues, we are left with just over 240 players to perform our data analysis on. Oftentimes, one of the hardest parts of a real-world data analysis is actually gathering and putting the data together. Now that we have a final data set to work with, we can go ahead.

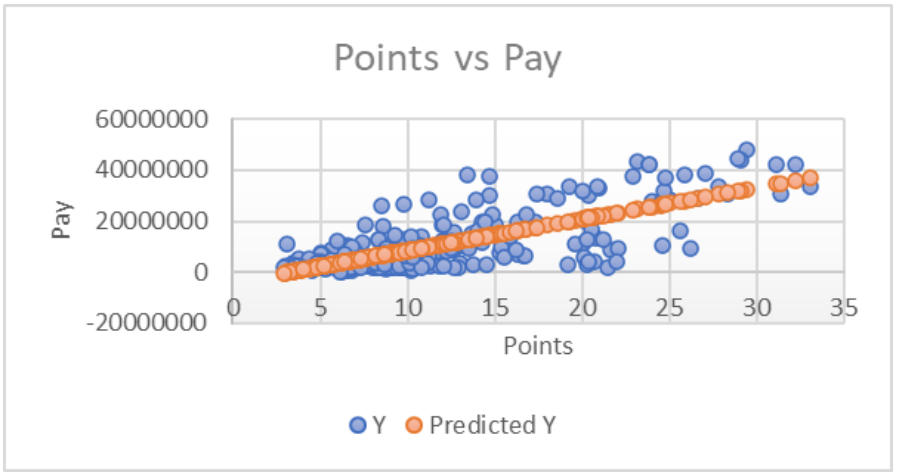
Here is a quick summary table of performance statistics.

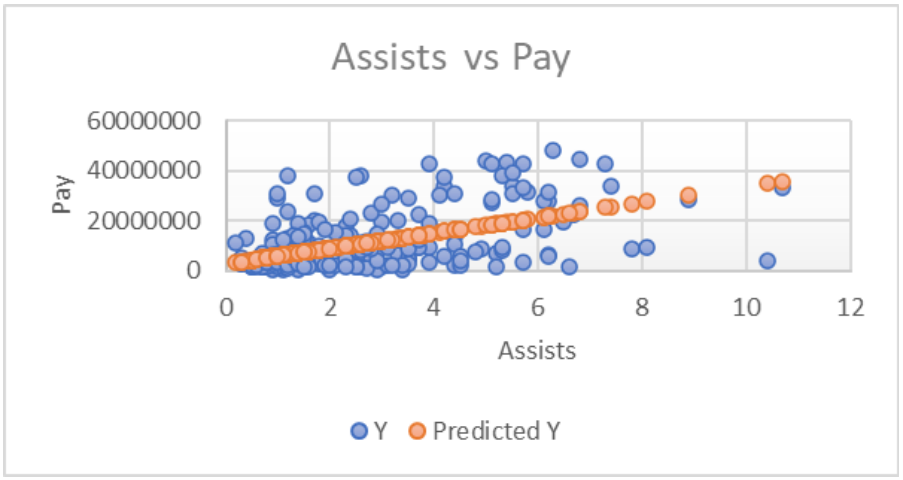
Summary Table (Per Game Basis)					
Stat	Points	Assists	Rebounds	Steals	Blocks
Mean	11.84	2.62	4.27	0.74	0.46
1st Quartile	6.6	1.1	2.7	0.5	0.2
Median	9.65	1.9	3.8	0.7	0.3
3rd Quartile	15.2	3.5	5.38	0.9	0.6

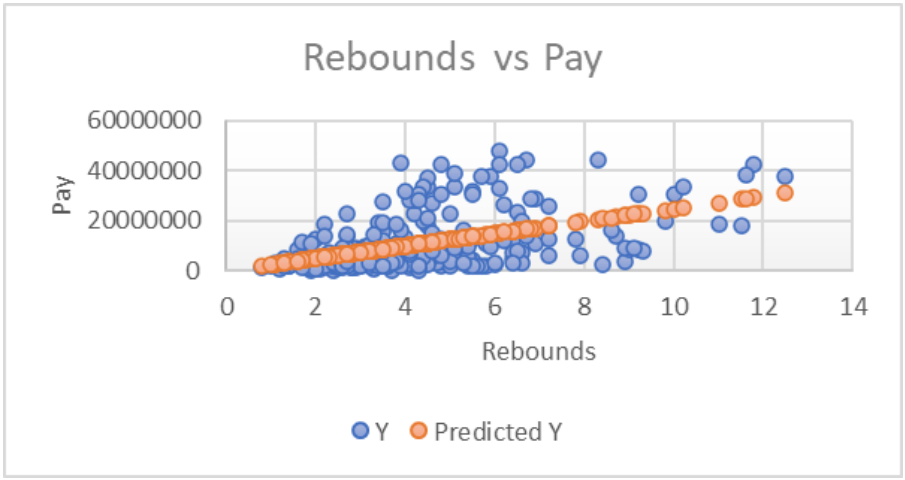
Immediately we can see that points per game offers the widest range of values. This intuitively makes sense as the objective of basketball is to score points and points are a multiple of baskets scored. The other stats recorded for the game are all subsidiary to scoring points or preventing the other team from scoring points. On a per game basis we may have a tendency to overvalue players who play a lot, and similarly undervalue players who don't get to play a lot of minutes. We can adjust for this later on by switching the counting method to a per 100 possession basis.

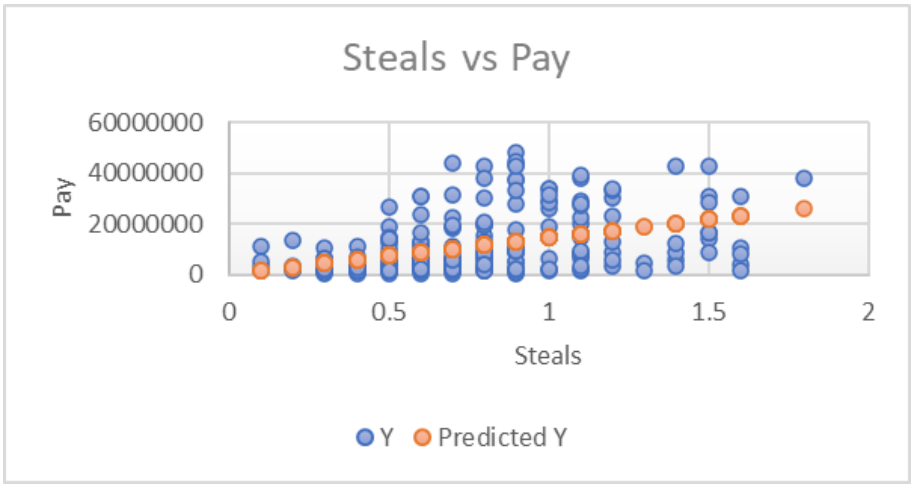
Now let's begin our first linear regressions. We'll start by just comparing these statistics to the pay of each player. A single variable linear regression will produce an estimate of what we would expect salary to be based on the closest fit linear line to the data. We can use the y-intercept and the slope of the line to create a plot which shows the predicted pay vs the actual pay of players. A couple of things to note with the regression analysis are some of the values that I will be referencing. The "Multiple R Value" provides an idea of the correlation between the compared data. A correlation coefficient can be between -1 (an increase in x will always result in a decrease in y) to +1 (an increase in x will always result in an increase in y), however the multiple r value provided in Excel's linear regression is always positive. The R-Square value gives us an idea of how well the data fits the prediction. R-Square ranges from 0 to 1, and a higher value indicates a better fit. One final statistic worth mentioning is the standard error. This value indicates how far off, on average, our prediction is away from the true values. It will be useful in this analysis because we can think of the standard error as the average dollar amount our prediction is off by.

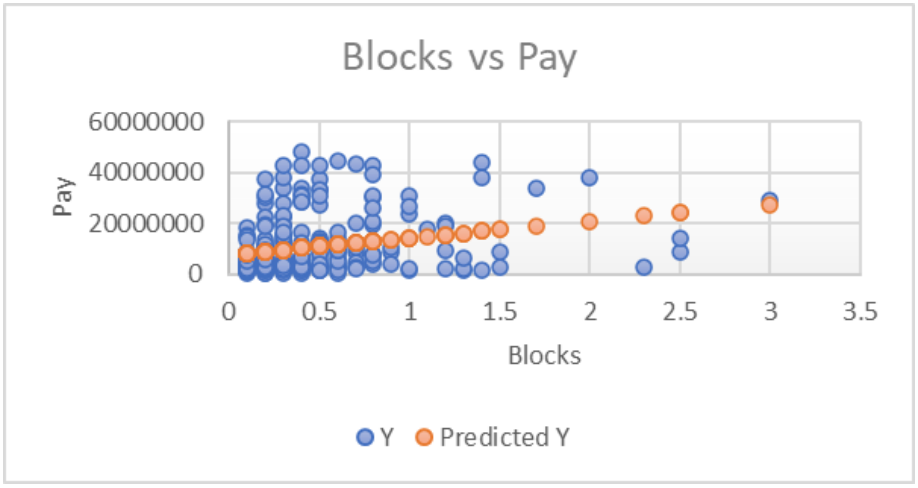
I think it's best to provide all 5 graphs and key regression statistics and then make observations afterwards vs going through each individually. This way we can avoid repeating information too often and only note what needs to be noted.


Points vs Pay
60000000
40000000
20000000
0
-20000000
Pay
0
5
10
15
20
25
30
35
Points
Y
Predicted Y


Assists vs Pay
60000000
40000000
20000000
0
Pay
0
2
4
6
8
10
12
Assists
Y
Predicted Y


Rebounds vs Pay
60000000
40000000
20000000
0
Pay
0
2
4
6
8
10
12
14
Rebounds
Y
Predicted Y





Summary Chart					
Stat	PPG	APG	RPG	SPG	BPG
Multiple R	0.747065545	0.531464602781001	0.483260951537253	0.432137244157183	0.249060969652743
R-Squared	0.558106928 52338	0.282454624009167	0.233541147280691	0.186742597787765	0.062031366 6043644
Standard Error	7622507.484 91865	9713236.170 4808	10038843.54 34173	10340779.45 72913	11105379.93 33365

It's immediately apparent PPG provides the best estimate of pay. It has the highest Multiple R value, the highest R-Squared value, and the lowest Standard Error. Points are usually thought of as the most important aspect of basketball as whichever team scores more points wins the game; it makes sense that we see the highest correlation with pay. Points also take on the

widest range of values which helps to smooth out the data and easier to approximate a line of best fit.

Also apparent, SPG and BPG seem to be bad indicators of pay. The R-Squared values associated with both of these stats are quite low. This makes some intuitive sense as these stats are usually considered among the least valuable of the 5 listed above. They also provide the smallest range of values they take meaning that outlier values are much more likely to mess with the fit of the line to the data. We can even see the range of values these stats take causing a much “bucketed” visualization of the data. This is important to keep in mind as we move forward to multivariate regression.

Now let’s look at the top 5 most overpaid players based on the predictions coming from each statistic’s linear regression. Because of the poor fit of SPG and BPG, I’ve decided to remove them from the following summary tables.

PPG	
Player	Overpay
1. Rudy Gobert	$25,515,752.73
2. Tobias Harris	$23,371,011.19
3. Draymond Green	$19,200,845.16
4. Bradley Beal	$18,520,511.88
5. Kyle Lowry	$18,393,465.50

APG	
Player	Overpay
1. Rudy Gobert	$31,754,247.02
2. Kawhi Leonard	$27,783,183.47
3. Anthony Davis	$27,263,442.58
4. Tobias Harris	$27,222,851.90
5. Kevin Durant	$26,032,663.98

RPG	
Player	Overpay
1. Bradley Beal	$33,501,259.01
2. Stephen Curry	$32,854,802.83
3. Damian Lillard	$30,490,183.66
4. Kevin Durant	$27,421,755.60
5. Paul George	$27,277,280.83

For the PPG top 5 overpaid players, what we observe in the top 5 generally makes sense. Most of the players included in the list have veteran experience and are usually thought of as defensive-minded players. Perhaps, the exception here is Bradley Beal who for most of his career has been a great scorer. Prior to the start of last season, Beal had just signed a large new contract paying among the league's highest salaries. At the same time, Beal's point production has dropped in recent years. Following 2 years of 30 plus PPG seasons, Beal has dropped to just over 23 PPG for each of the last 2 seasons.

Similarly, what we see for the APG top 5 overpaid players seems to pass the eye test as well. Rudy Gobert again tops the list, but as a defensive minded center, his talents have never been in dishing the ball around. It just so happens that these regressions do not look favorably on any production Gobert is providing his team based on these selected stats. This might indicate that Gobert is indeed a highly overpaid player, but let's dig deeper before rushing to any conclusions.

RPG also passes our common sense test, but with a different set of players this time. We would expect positions like centers and power forwards to get more of the rebound share on their team. Subsequently, because of position and play, we expect to see smaller guards and small forwards rebounding less. This regression thus punishes players who have large contracts, but whether it be position, size or both, do not get as many rebounds.

What if we were to take the average rank of each player in this overpay analysis and see what happens? In doing so, we would be weighting each category equivalently which is probably incorrect because we know that points are how a team wins games. But, it's a starting point and we can adjust our formula after we have obtained these values. Although, it might be difficult to do this; we know that points are almost certainly more valuable than rebounds, but assigning a numerical weight for the value of one statistic over another would be subjective.

This is the table of 5 highest average positions in the regressions from PPG, APG, and RPG.

Average Position (Based on Overpay)					
Rk	Player	PPG	APG	RPG	Average Position
1	Bradley Beal	4	9	1	4.666666667
T2	Tobias Harris	2	4	10	5.333333333
T2	Kawhi Leonard	8	2	6	5.333333333
4	Stephen Curry	10	6	2	6
5	Paul George	7	8	5	6.666666667

This table shows us, based on the methodology used for this part, that Bradley Beal is the most overpaid player in the NBA for the last season. This doesn't seem to be a wildly inaccurate assertion based on some lists online. There aren't actually all that many comprehensive lists and methods posted online which attempt to answer this question in a rigorous and quantitative manner. In the list referenced above, the calculation is done based on a statistic called "real value" which takes into account many things that are beyond the scope of what will be done for this analysis. For one, the description of the statistic references that players *off court* issues can potentially impact their value. The statistic also takes in playoff performance, injuries, and multiple seasons to make their prediction of value.

Let's move on to multivariate regression. For this multivariate regression, we will be using PPG, APG, and RPG as our x-variables. We know that these values gave us our highest level of predictive power of the 5 stats we examined above. Now that we are doing multivariate regression, we will no longer be able to visualize the data with a simple x-y graph because we have added more "dimensions" to our prediction.

Summary Performance of Multivariate Regression	
Multiple R	0.762661683
R-Squared	0.581652843
Standard Error	7447745.3132437

We see that the accuracy of this regression has improved. Not by a drastic measure, but it's good to see that we have our highest observed Multiple R, R-Squared, and our lowest Standard

Error. It's an indication that we are on the correct track. Now, let's observe the 5 most overpaid players based on this regression.

Multivariate Regression Overpay (PPG, APG, RPG)	
Player	Overpay
1. Tobias Harris	$22,923,000.01
2. Rudy Gobert	$20,390,391.46
3. Bradley Beal	$20,032,668.70
4. Kawhi Leonard	$17,306,663.25
5. Paul George	$16,923,868.99

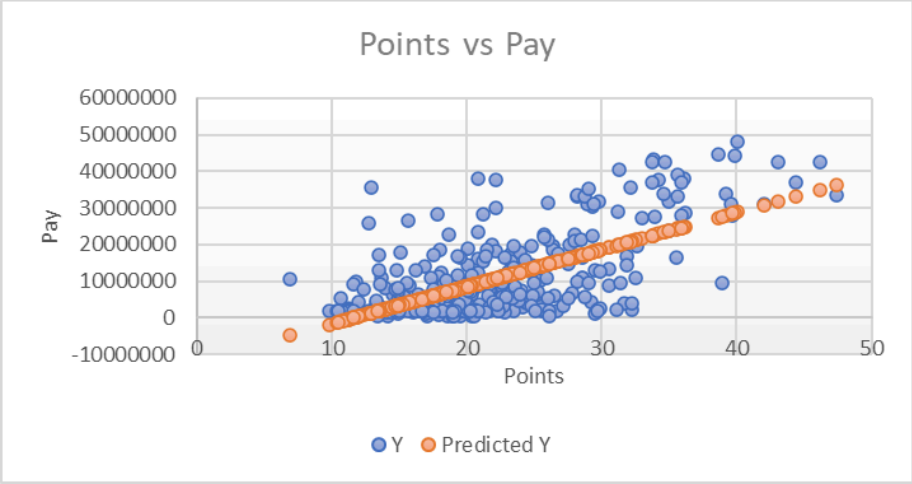
We can see a couple of interesting things in this table. Firstly, I would like to point out that the Clippers have two players in the top 5 most overpaid players in the league based on this regression. That's alarming for Clippers fans as it suggests that Kawhi and George are already under performing, and given their ages (32 and 33 respectively) are quickly aging out of their primes. We again see Bradley Beal and Tobias Harris topping the charts, which makes sense, given the similarities of this multivariate regression to the single regressions we ran earlier. However, we see a new entrant as the second most overpaid player in the NBA: Rudy Gobert. Gobert was the most overpaid player based on both of the first two single regressions run (PPG and APG). But, in the RPG analysis Gobert's rank in overpay dropped all the way to 38th, with the 1st position being the highest overpay. This was enough to keep him out of the top 5 based on the average position of all three regressions shown in the earlier table.

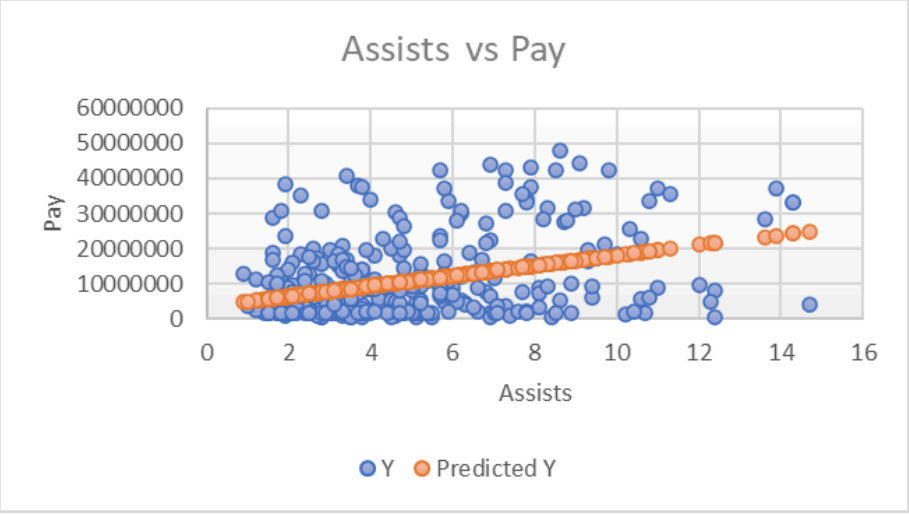
Let's now move to linear regressions done with statistics done per 100 possessions. Shifting the statistics to this method averaging allows us to gain insights into players on an equal playing field. It allows us to cancel the effect of varying play time among players in a game. However, it's not without drawback. It can sometimes have the effect of elevating the statistical impact of bench players because it gives them an edge in the ways the statistics are recorded. If a bench player comes off the bench for a few possessions, their statistics recorded over 100 possessions could reflect performance that would be impossible for them to produce over any sustained period of time. Per game data can highlight the performance of the starting players, but this could be a benefit to the performance analysis relative to pay we are conducting. Nevertheless, it will still be a useful continuation of the analysis done above to rerun our regressions of points, assists, and rebounds using a per 100 possessions basis.

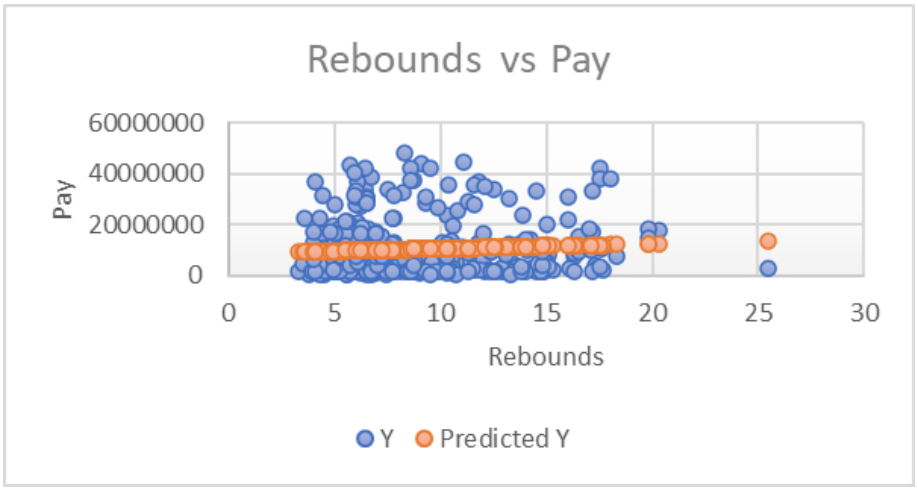
This is what the performance summary looks like with a per 100 possessions method of recording the statistics.

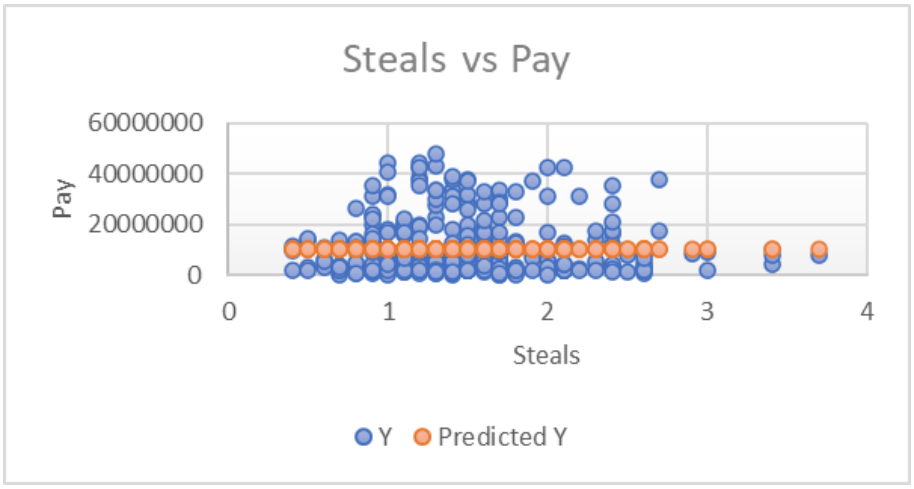
Summary Table (Per 100 possessions)					
Stat	Points	Assist	Rebounds	Steals	Blocks
Mean	21.90	4.74	8.78	1.45	0.97
1st Quartile	17.18	2.6	5.8	1.1	0.4
Median	20.5	3.9	7.5	1.4	0.7
3rd Quartile	25.7	6.125	10.85	1.7	1.2

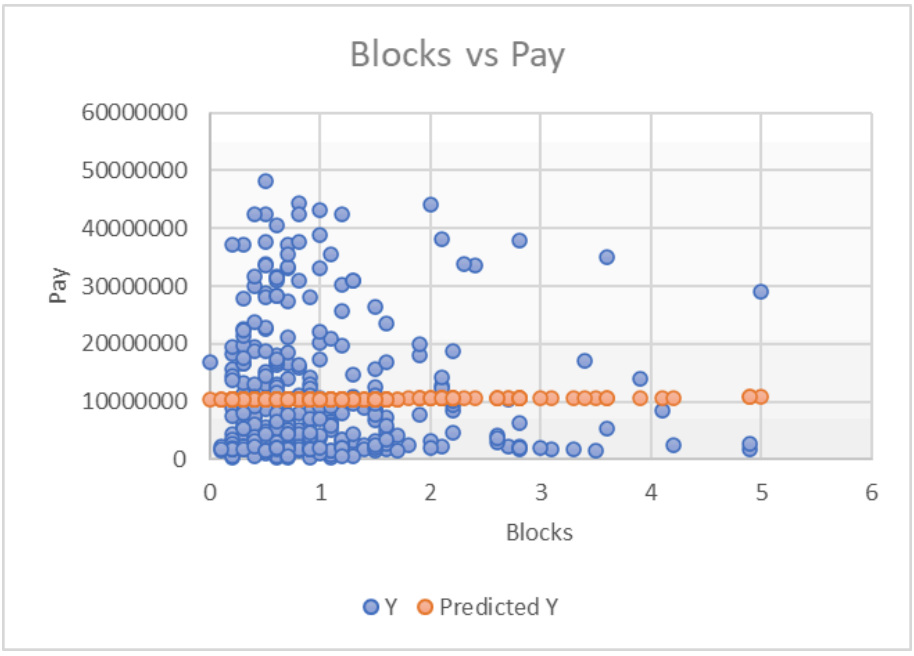
Regression outputs using the per 100 possessions methods look like this.






Rebounds vs Pay
60000000
40000000
20000000
0
Pay
0
5
10
15
20
25
30
Rebounds
Y
Predicted Y


Steals vs Pay
60000000
40000000
20000000
0
Pay
0
1
2
3
4
Steals
Y
Predicted Y


Blocks vs Pay
60000000
50000000
40000000
30000000
20000000
10000000
0
Pay
0
1
2
3
4
5
6
Blocks
Y
Predicted Y

Summary Chart					
Stat	PP100	AP100	RP100	SP100	BP100
Multiple R	0.647384242	0.362523259	0.067741077	0.001386663	0.006364609
R-Squared	0.419106356	0.131423113	0.004588854	1.92283E-06	4.05082E-05
Standard Error	8517683.447	10415432.72	11149988.86	11175649.33	11175433.72

Every single regression using the per 100 counting method returned worse results than our original per game regression. That is to say, the predictive power of our regressions decreased for each statistic. It seems that the issues mentioned above indeed negatively impacted our model. The steals and blocks regressions returned almost hilariously bad results.

I now want to take some time to discuss some of the issues with conducting a linear regression on pay. Pay in the NBA, by definition, isn't linear. There is both a maximum a player can earn, and a minimum. A further complexity to this idea is that maximum pay is different for players. A player that is resigning with his team that drafted him can earn more than if he signs with another team. Rookie contracts are capped based on draft position, so we see players that massively over perform their contracts but aren't eligible for more compensation. Further analysis using a non-linear regression might be interesting to perform. In this type of regression, a non-linear line of best fit is drawn through the data to give a better idea of how players perform relative to pay.

The market for NBA players is relatively illiquid. Even among pro sports, the NBA rosters the fewest players per team, so there are relatively fewer options (players) for a team to select from. This may force teams to overpay for some of their players simply because they lack a viable alternative, and other teams may be willing to shell out to get performance from players.

The contracts are biased to pay more to whoever has signed the latest contract. Maximum contracts are based on a percentage of salary cap, which usually expands every year. This phenomenon has led to players like Mike Conley signing the largest contract in NBA history (at least for a time) despite not really being in the conversation of league's best player.

Appendix

1. https://www.basketball-reference.com/leagues/NBA_2023_per_game.html
2. (PPG, Points per Game), (APG, Assists per Game), (RPG, Rebounds per Game), (SPG, Steals per Game), (BPG, Blocks per Game)
3. https://www.espn.com/nba/salaries/_/year/2023/seasontype/1
4. https://www.statology.org/multiple-r-vs-r-squared/
5. https://hoopshype.com/lists/most-overpaid-players-nba-2022-23-towns-gobert-westbrook-simmons/
6. https://hoopshype.com/lists/how-hoopshypes-real-value-works/