Balling on a Budget: An Analysis of Player Performance and Salary in the NBA

Charles Reynolds, Nandini Rajgopal, Emma Edelman, Stephanie Tungerapan, Pierce Renio

Introduction

This report is focused on identifying statistical trends between the 2018 and 2019 seasons of the National Basketball Association (NBA). Moreover, the purpose is to identify potential variables that influence player salary and analyze its significance within the framework of professional athletics. Analyzing data on salary and its association with player performance is an important tool that NBA organizations look into to determine their payroll efficiency. Each NBA team has limits set by the league on how much money in total they can pay out in salary and so it is in their best interest to use that money to pay the best performing players so they will remain on the team. Analysis of salary across all players in the league will allow interpretations on what variables are most associated with salary and so decisions can be made on what a player's expected pay would be based on their statistics. Some factors have more prevalent correlations than others, so it's important to provide a wide context from which the data can be properly extracted and computed.

The data in question includes two datasets from both the 2018 and 2019 NBA seasons respectively. Each dataset shows the complete roster of all active NBA players along with their corresponding statistics which include but are not limited to: team, height, weight, salary, age, games played, and average points. There are two main questions to be answered from the data: how has player salary changed between the seasons, and what possible factors may lead to this result. We hypothesize that statistics that best analyze player performance have a larger association with salary than those statistics that are less wholistically evaluative. To do this we first look at broad trends within the data, and then narrow our own search. For the report, we are specifically looking at general salary alongside variables such as average points per game, team, and total games played within a season.

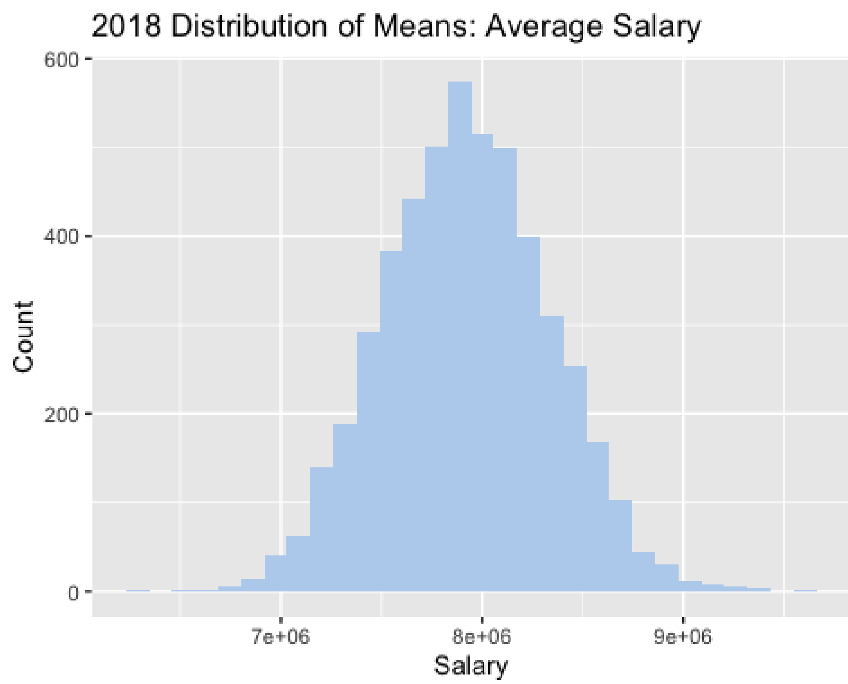
Due to the wide discrepancy between the number of players per season and the total games played, we thought it sensible to only look at data from players who played at least 41 games (or 50% of the maximum games played within a season). This data is more representative of players who are able to participate full time. In addition, narrowing the data allows the elimination of potential outliers.

Exploratory Data Analysis/ Description of Data

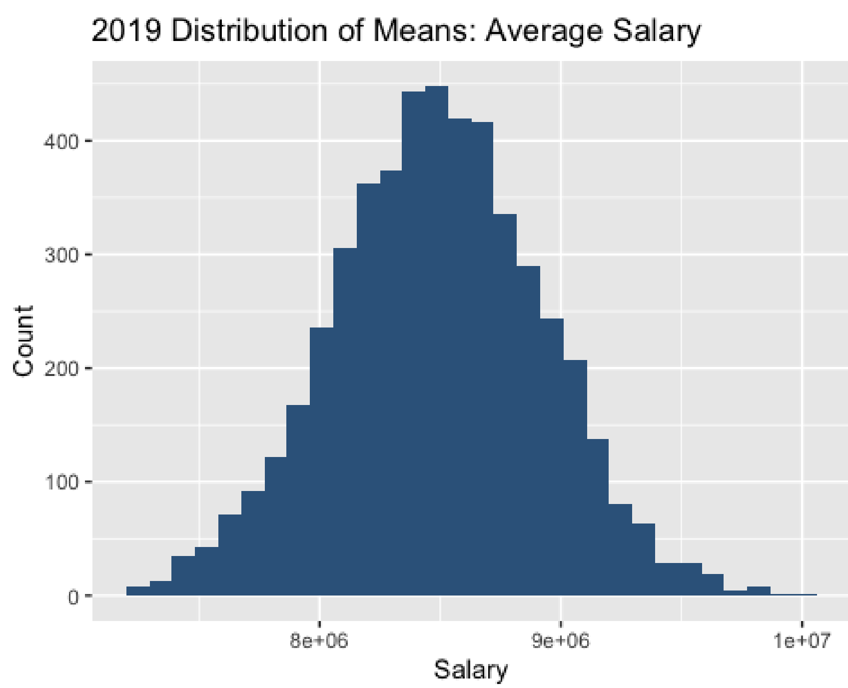
Salary Histogram - Charles Reynolds

The first statistic that we analyze is player salary, in order to get an understanding of the spread of salary from year to year. To do this, we plot the salary of each player along a histogram for each season. By doing this, we can easily look at the general trend of player income between seasons. In order to get an accurate grasp of player salary, it is helpful to normalize the data to fall under a traditional bell curve. As such, the distribution of average salary is plotted by taking a random sample of 350 and replicating it by 5000. The result is a normalized bell curve that can later be looked at with hypothesis testing.

```
## [1] 7914162
```



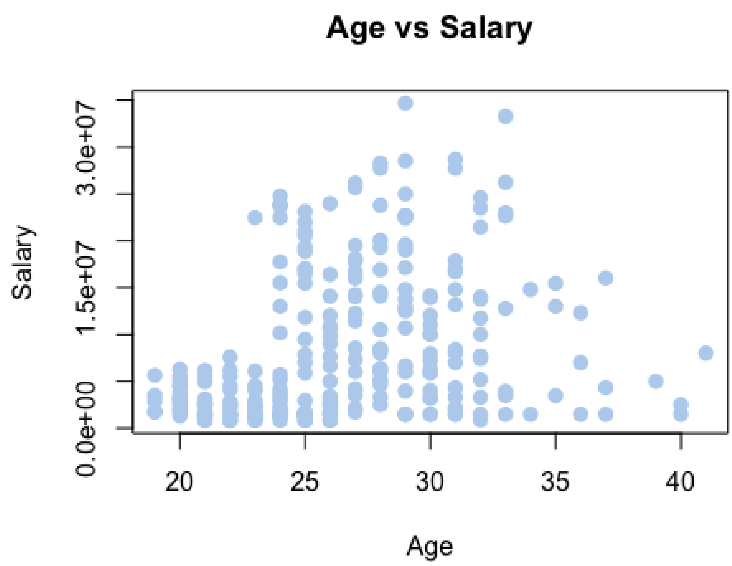

```
## [1] 8469914
```

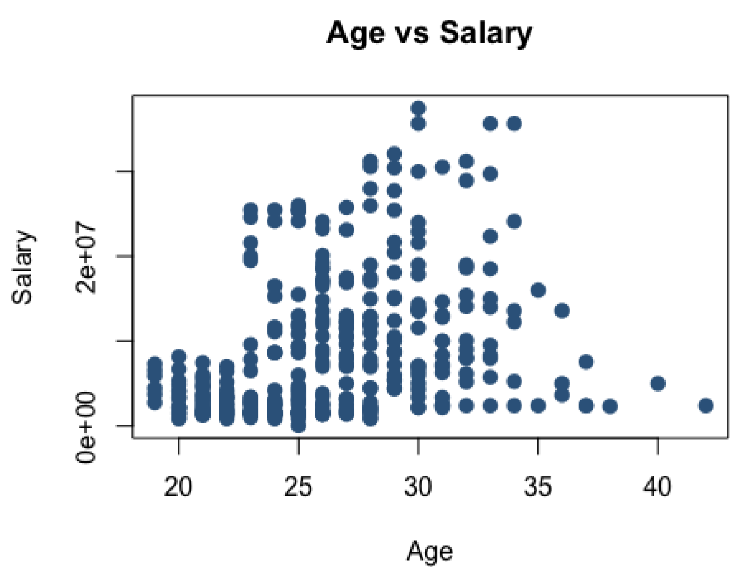


As we can see from both histograms, the plots are normally distributed over the average salary between the two seasons. It's important to note the similarity in the distribution of salaries. Even though both graphs are normally distributed, the 2019 average appears to be larger than the 2018 average, thus possibly suggesting an increase in pay between seasons. We then move to try and find what possible factors could be causing or not causing the change between each season's salary.

Age and Salary - Stephanie Tunngerapan

Now that we have a better understanding about the distribution of salaries, we want to narrow our analysis and compare salary to a variable that could possibly influence it. For that we analyze and plot the relationship between player ages and their salaries.

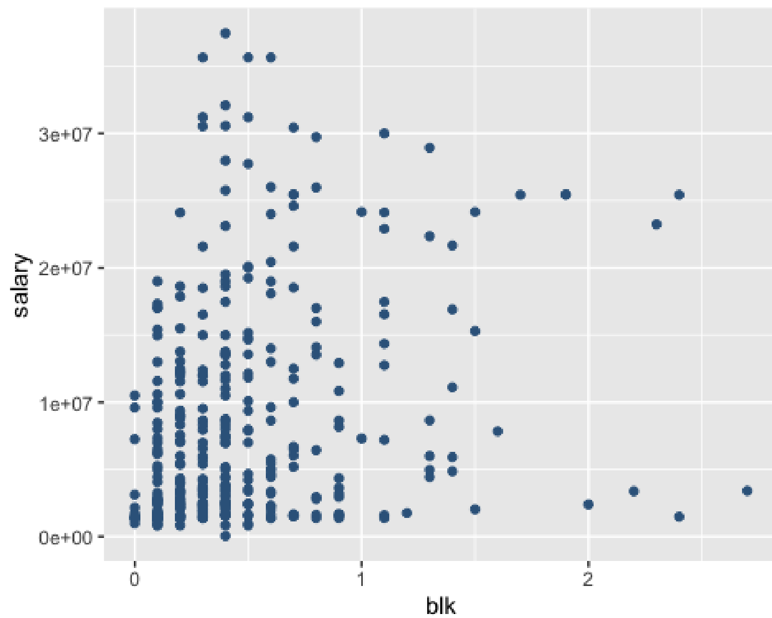


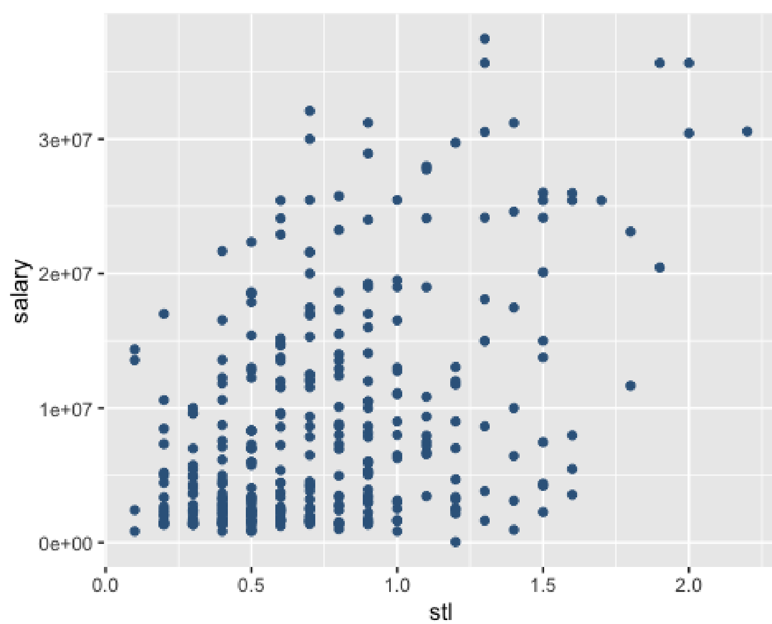


By looking at the two plots, we can see that the distribution of salaries over age is quite similar between the 2018 and 2019 seasons. Most players in the NBA are typically between the ages of 25 - 35. In addition, we see a clear distribution of salaries, which appears to be almost normally distributed. Within the 25 - 35 age range, players seem to be getting paid the highest salaries, and those with the lowest salaries fall outside the range. Thus, age does play some role in determining player salary.

Steals and Blocks - Pierce Renio

To continue to find potential explanatory variables for salary, we analyze specific statistics that evaluate player performance. To start, we look at the average number of blocks and steals compared to player salary.

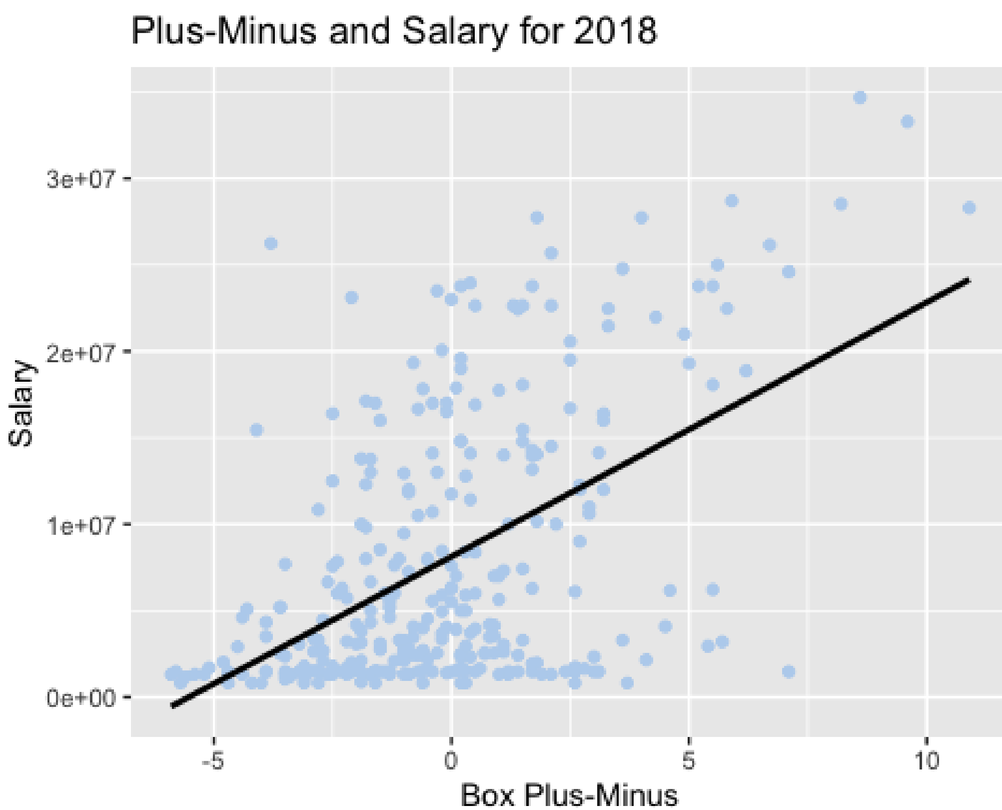




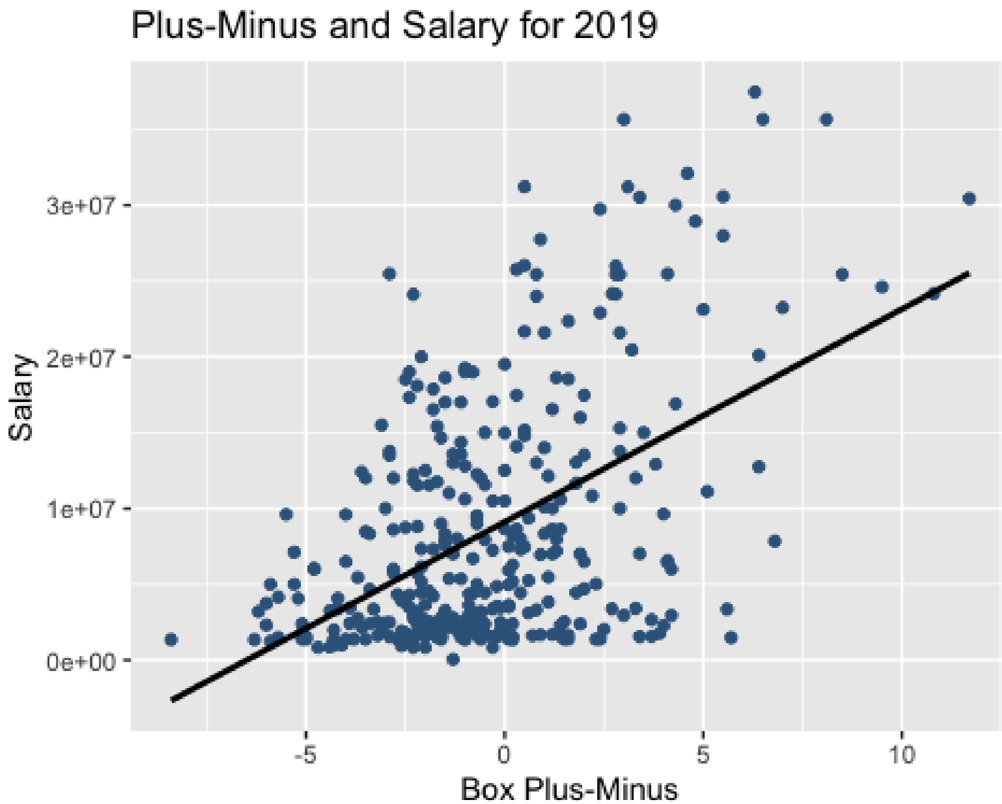
Because the 2018 data for block and steals was rounded, the best analysis came from the 2019 data to make conclusions. As is evident, most players have a low number of average blocks and steals. There is a very low positive correlation between salary and steals. This data helps provide a better understanding about the distribution of salaries, but it does not provide substantial insight on to the determination of salary. To continue, we look to more holistic player performance statistics to try and find an association with salary.

Plus-Minus and Salary - Emma Edelman

To continue our exploration into player performance and salary, we look at a different statistic that provides a more complete evaluation of player performance, Plus Minus. Plus Minus has gained traction in data analytics for the NBA as it provides a valuable estimate of a player's contribution to a team. This statistic calculates the value a player brings to the team by evaluating the score of the game while the player is in. For our analysis, we used a slightly more complex version of this statistic: Box Plus Minus. This is a rate statistic that calculates player Plus Minus as an average per 100 possessions, and then compares it to league average of 0.0. This statistic therefore provides a more complete evaluation of player performance and we thus hypothesize it will have a stronger association with salary.




```
## [1] 0.5240336
```

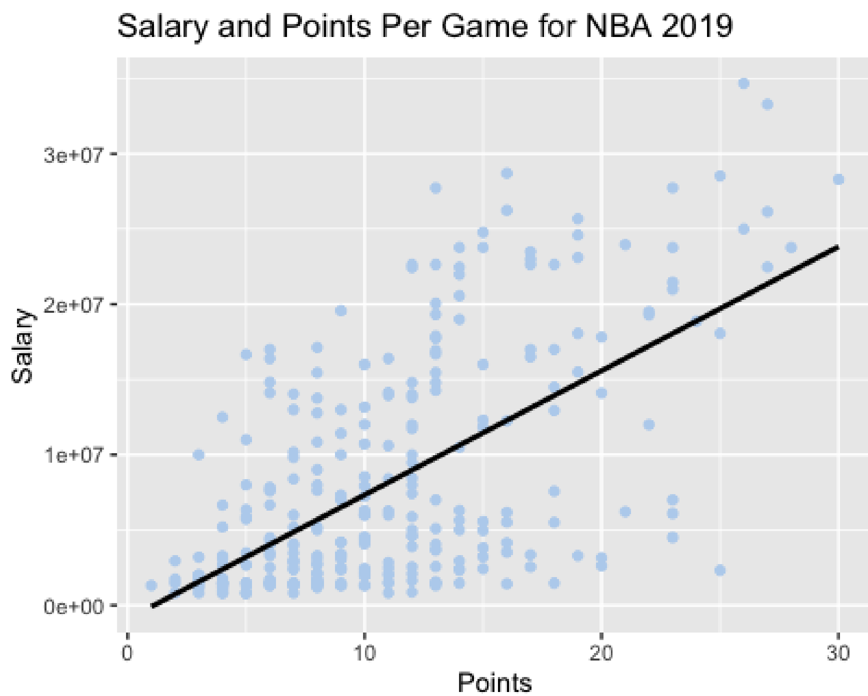


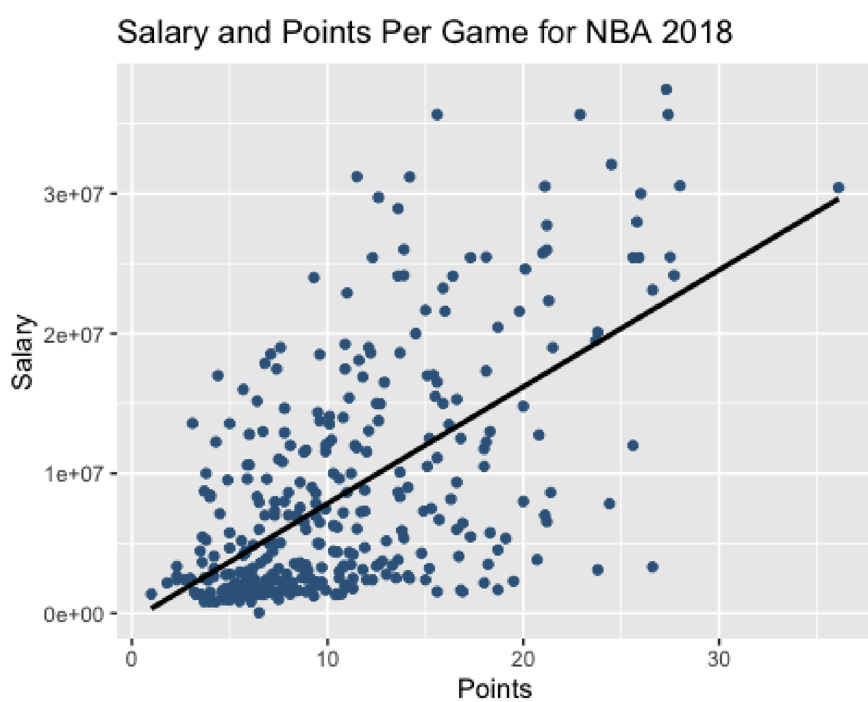

```
## [1] 0.5007454
```

Plus Minus measures an individual player's performance compared to players in the rest of the league, so we expect it will have some association with salary. After plotting a scatter plot of players' Plus Minus and salary we find a positive correlation of around 0.5 between salary and Plus Minus for both seasons. There is almost no change between seasons and the linear regression lines for both seasons have very similar slopes and intercepts, implying that Plus Minus has a constant positive association with salary from year to year. This stronger association with a more evaluative player performance statistics, Plus Minus, seems to follow our hypothesis.

Points and Salary - Nandini Rajgopal

To continue or analysis of how player performance may impact salary, we look at arguably the most obvious player performance statistic: points. A player's performance is quite often compared to the amount of points they make, so it is useful to understand the correlation between a player's average points per game compared to his income.



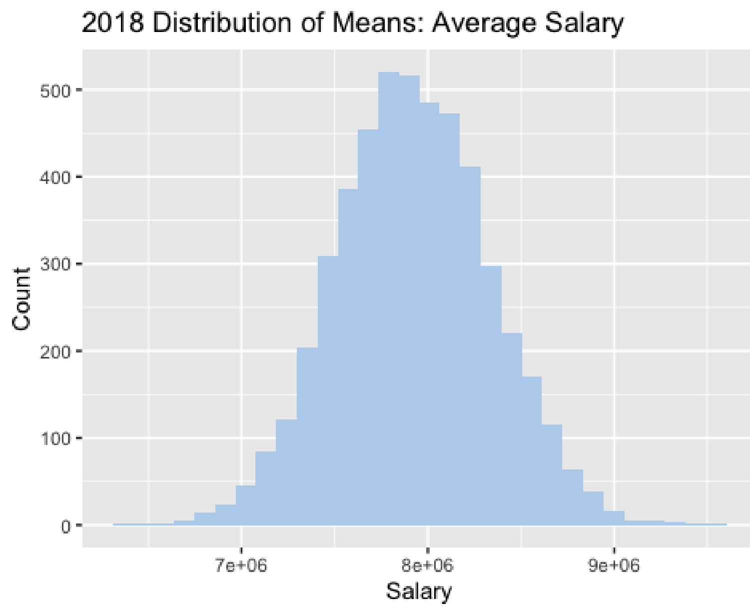


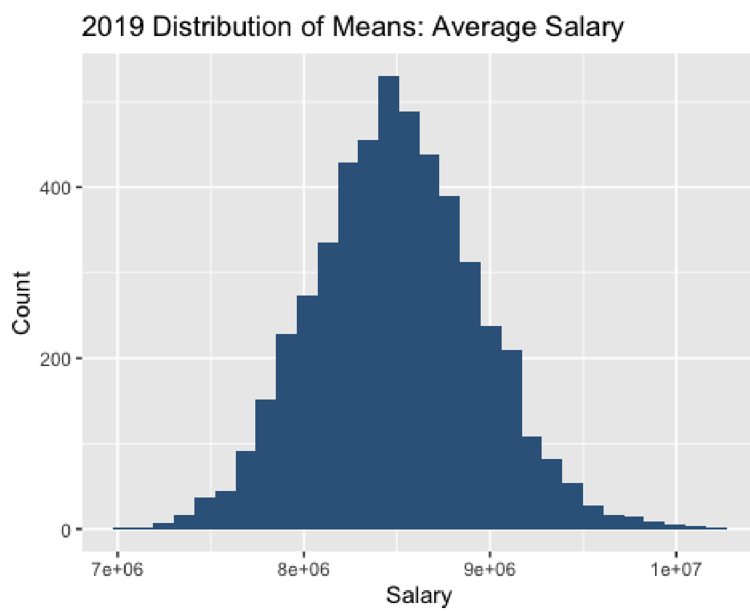
As we can see from the 2019 scatterplot, the average points scored is highly concentrated and skewed to the left. We also get a better view of the concentration of salaries in the 2019 season as compared to 2018. The correlation between salary and points in 2018 appears to be positive and linear, at least more so than the 2019 data. In general, there appears to be a larger distribution of salaries in 2019, meaning the range of potential earnings is larger in 2019 than in 2018. In regards to points, the distribution of points appears to be similar between seasons.

Data Analysis

Of all the data that has been plotted and computed, there is an overarching question that must be answered first: did players' salaries actually change between seasons? To do this, we look at our salary histograms and utilize the double-sample z test on the normalized distribution of salary averages between seasons. Our null hypothesis is that the average salary is the same between both seasons, and our alternative hypothesis states that the two averages are actually statistically different (or something else). In other words, we are investigating whether the average salary actually did increase in the 2019 season.

Salary Between the Two Seasons



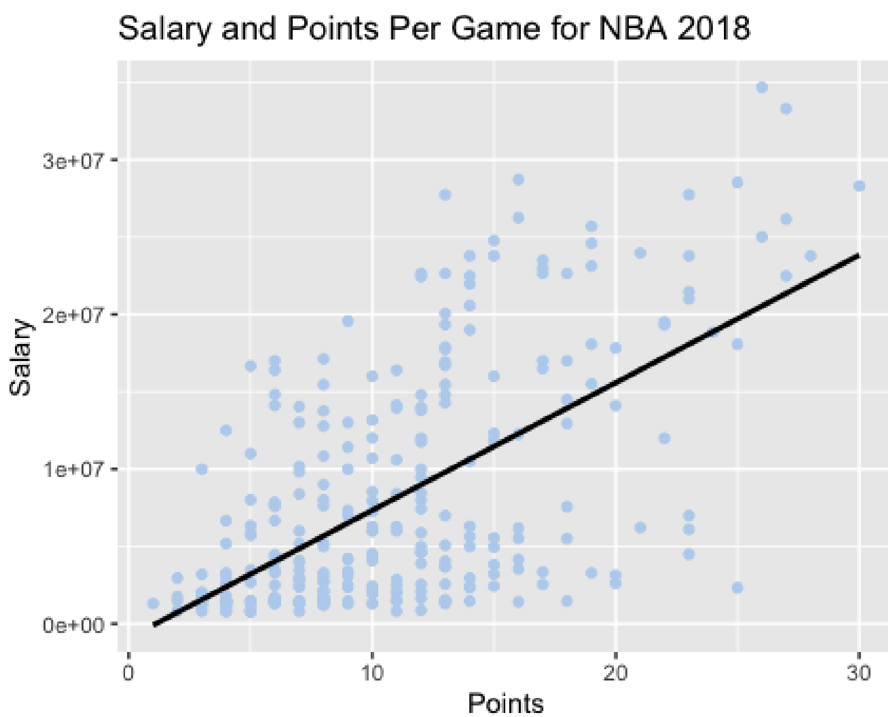



```
## [1] 5.259781e-09
```

The resultant of the test is a p-value that is approximately 0%, thus we can reject the null hypothesis and accept the alternative. So based on the results, there does appear to be a statistically significant difference between the distribution of salary averages between seasons. In other words, given two random samples of players from each season, the players from the 2019 season are likely to collectively have (on average) a higher salary. The NBA as a league collectively saw players being paid more, though this does not suggest that all players got a pay raise.

Points and Salary

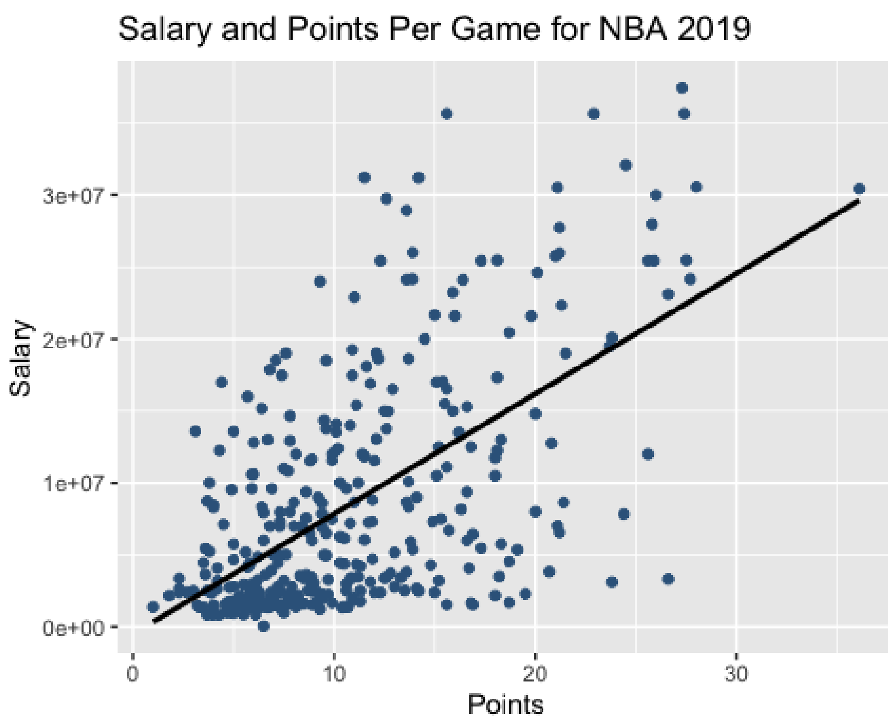
Now that we have shown there was a change in salary, we look to analyze what statistics could explain this shift. We chose to take a closer look at the last variable we explored in our Exploratory Data Analysis– points. There seems to be a positive correlation between salary and points, so we hypothesize that there will be an increase in points per game, correlated with the increase in salary. We then proceed to investigate this correlation between points and salary, to conclusively determine whether or not there was a relationship.




```
## [1] 0.6180728

##
## Call:
## lm(formula = salary ~ pts, data = nba18)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17388529  -4030984   -912441   3469308  17925982
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -915999     738832   -1.24    0.216
## pts           824956      60579   13.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6145000 on 300 degrees of freedom
## Multiple R-squared:  0.382,  Adjusted R-squared:   0.38
## F-statistic: 185.4 on 1 and 300 DF,  p-value: < 2.2e-16
```




```
## [1] 0.6012156
```

```
##
## Call:
## lm(formula = salary ~ pts, data = nba19)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -18390320  -4101742  -1405926   3787038  23128469
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -491816     730748  -0.673    0.501
## pts           834455      59453  14.035   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

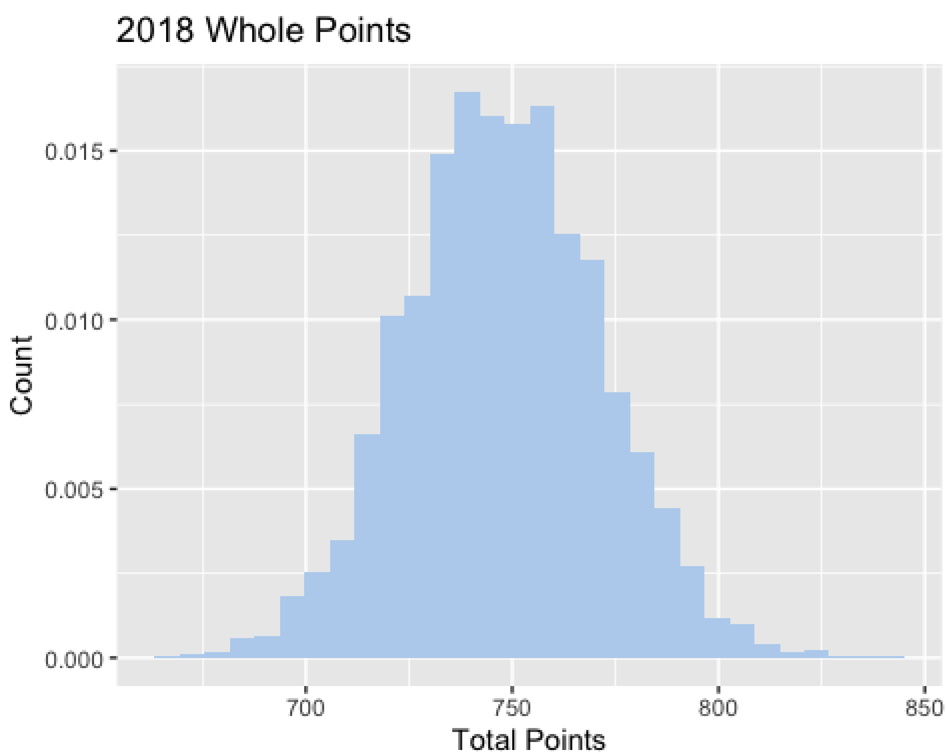
```
##
## Residual standard error: 6608000 on 348 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3615, Adjusted R-squared:  0.3596
## F-statistic:   197 on 1 and 348 DF,  p-value: < 2.2e-16
```

From this linear regression analysis, we can determine that the correlation between points and salary in both 2018 and 2019 was around 0.6. This correlation is positive and somewhat strong, suggesting the evidence of a positive relationship between points and salary. In addition, while the regression lines are similar, the slope of the line in 2019 is larger than the line in 2018. This suggests that the higher scoring players scored even more points in 2019.

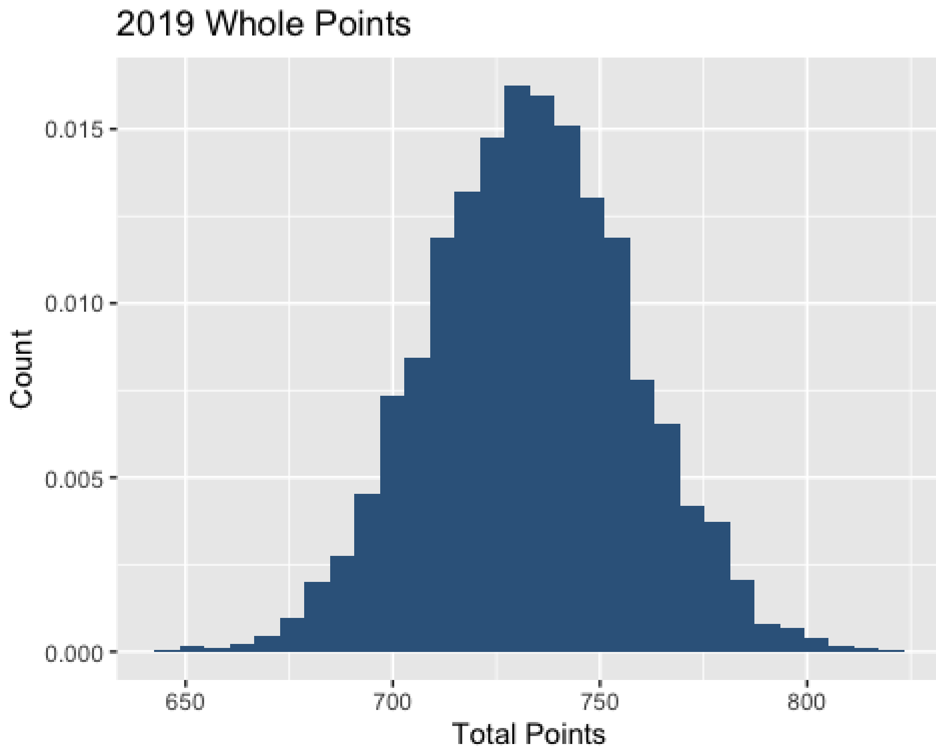
Total Points

The linear regression analysis got us to postulate that because there is such a strong relationship between points and salary, there may be an increase in total points between the two years. In order to effectively analyze this difference in total points, we create our own variable by multiplying points and games played. This creates a variable of total points per player in each NBA season. We then compare the total points between each season in both a histogram and z-test. Our null hypothesis for the z-test attributes the difference in points to chance, and argues that there is no real difference between the seasons, while our alternative hypothesis deems this difference significant, and not due to chance.

```
## [1] 747.8494
```




```
## [1] 732.9737
```




```
## [1] 99.89679
```

After comparing histograms with the total points of players between the two seasons, the average of the total points for 2019 at around 524 is noted to be higher than the average in 2019 at 463. We then perform a z-test to determine the significance of the difference between the averages. Upon completing this hypothesis test, the p-value is determined to be very close to 0. Therefore, we reject the null hypothesis, and assume the difference between the two seasons to be significant. The average of the total points increased between the two years. This supports our evidence of a strong relationship between points of an NBA player and their salary.

Conclusion

As evident in our hypothesis tests, we identified numerous interesting trends. From the distribution of means, we computed the spread of averages from created samples. According to the data, the change in the distribution of average salary is significant and not due to chance. In addition, we concluded that the average total points per player for each season collectively increased in 2019. As a whole, we identified a positive trend between a player's points and their salary.

Statistically speaking, players who score more points get paid more. But even though this may be the case, the correlation is not perfect and there are plenty of other examples suggesting otherwise. This may be because the NBA is very competitive, and as the league continues to evolve there is less of a talent gap abroad. In other words, the NBA now has more depth than ever before and players are more competitive, meaning there is less incentive to pay players based on performance alone. In other words, performance is evidently quite important in how a player gets paid, but it is by no means the only decisive factor. This may illustrate a tendency to pay players based on marketability and brand potential rather than actual quality of performance.

It's important to note that the data only represents the NBA as a league and doesn't take into account individual player performance and improvement. The data illustrates broad trends within the context of the entire franchise, meaning there are discrepancies amongst individual players that could create bias. Such discrepancies may include but are not limited to: injuries, player position, overall team strength, facilities/financial resources, team location, fanbase etc. In addition, correlation does not necessarily equate to causation, meaning the trends that have been analyzed may not apply to all players.

References

1.2017-18 NBA Player Stats: Totals. (n.d.). Retrieved from https://www.basketball-reference.com/leagues/NBA_2018_totals.html.

2.2018-19 NBA Player Stats: Totals. (n.d.). Retrieved from https://www.basketball-reference.com/leagues/NBA_2019_totals.html.