Final Project

due March, 28, 2022 by 11:59 PM

Jiewei Li, Taylor Glatt, Ayush Jain, Juliette Clark, Matthew Paul

March 27, 2022

Introduction and Data

Being Duke students in the middle of March Madness season, we are constantly bombarded by basketball coverage. Whether it be Chuck on NBA on TNT or Turner's March Madness coverage, we often see several statistics and hear opinions based on these statistics. However, it would be interesting to unravel the statistics behind the sport and make our own analyses based on our findings. For our topic, we will be using data from FiveThirtyEight called nba-raptor. RAPTOR is FiveThirtyEight's in-house NBA statistic which uses NBA data to calculate statistics. The data was collected in 2021 and includes data going all the way back to 2014. However, we will only be using data from 2018. Since our research topic is the NBA, the research question we decided to investigate was:

Are there meaningful correlations between minutes played and offensive scores / defensive scores during the 2018 season?

Further, we will also compare whether or not these conclusions are the same if the scores are taken from the regular season (RS) or the playoffs (PO).

We hypothesize that the players with the most minutes played will have the highest raptor offensive and defensive scores. Further, we hypothesize that there will be no meaningful difference between the mean offensive / defensive scores of the regular season and the playoffs.

Some relevant variables we will be using in our analysis include:

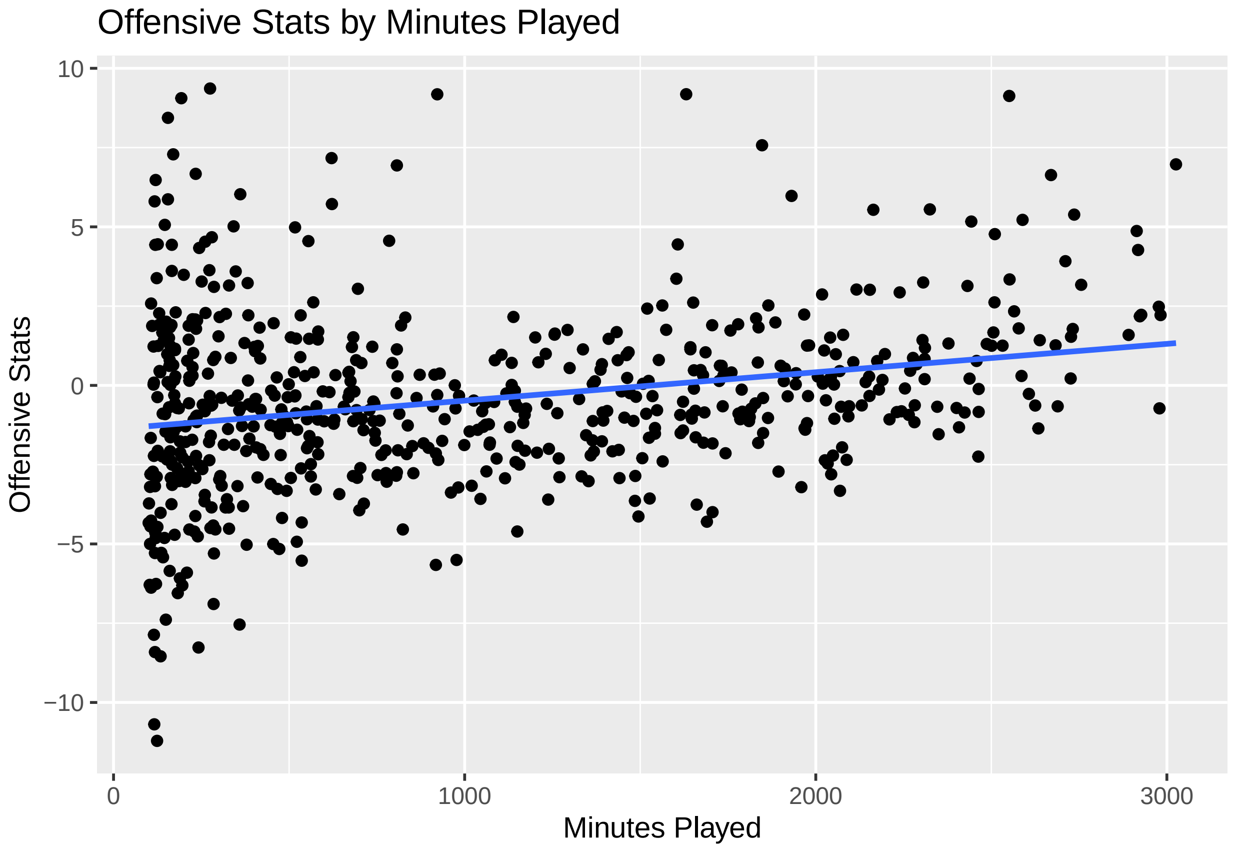
1. player_name: this variable contains the name of the player
2. season_type: this variable has values of either 'RS' or 'PO' indicating whether the row contains the player's regular season statistics or playoff statistics
3. mp: this variable indicates the minutes played by a player. For the purpose of our research question, we will only be looking at players with over 100 minutes played (equivalent to roughly two NBA games)
4. poss: this variable indicates the number of possessions played by the player
5. raptor_offense: this variable calculates the RAPTOR offensive score for a player based on offensive statistics. Negative values indicate bad performance, positive values indicate good performance
6. raptor_defense: this variable calculates the RAPTOR defensive score for a player based on defensive statistics. Negative values indicate bad performance, positive values indicate good performance
7. raptor_total: this variable sums up the player's RAPTOR offensive and defensive scores

Methodology

Part 1

What are the correlations between minutes played and offensive statistics? Defensive statistics?

```
## `geom_smooth()` using formula 'y ~ x'
```



```
## # A tibble: 2 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -1.38      0.174        -7.90 1.26e-14
## 2 mp           0.000895  0.000135      6.62 7.67e-11
```

```
## [1] 0.06455444
```

$$\widehat{RaptorOffense} = -1.38 + 0.0009\ MinutesPlayed$$

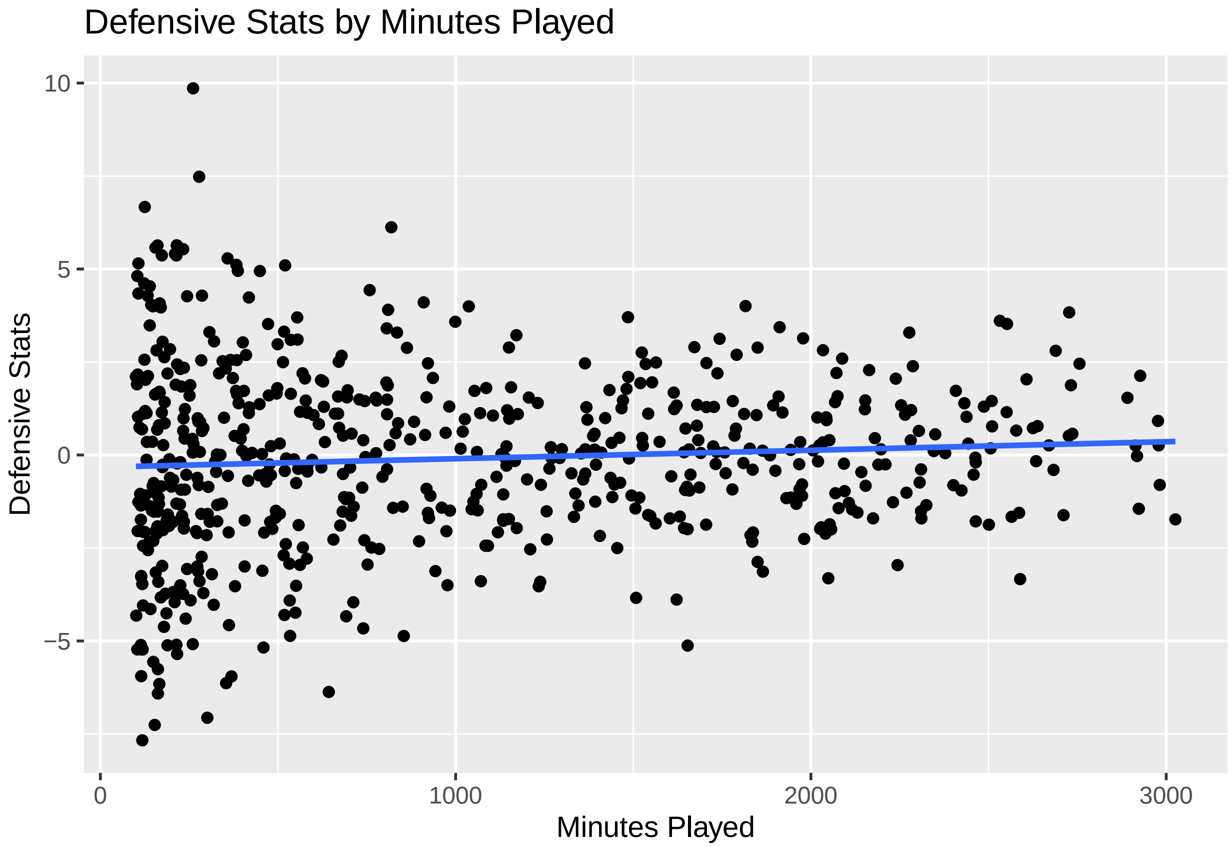
Intercept: A player without a single minute of playing time, would have an offensive score of -1.98. This is meaningless in context.

Slope: A one minute increase of playing time would predict a 0.0009 percent offensive score increase.

This data indicates that there is a slight correlation between playing time and good offensive stats.

The adjusted r-squared value of 0.065 is low and indicates weak effect size.

```
## `geom_smooth()` using formula 'y ~ x'
```



```
## # A tibble: 2 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) -0.328     0.155        -2.11  0.0353
## 2 mp           0.000228  0.000121      1.89  0.0589
```

```
## [1] 0.004137558
```

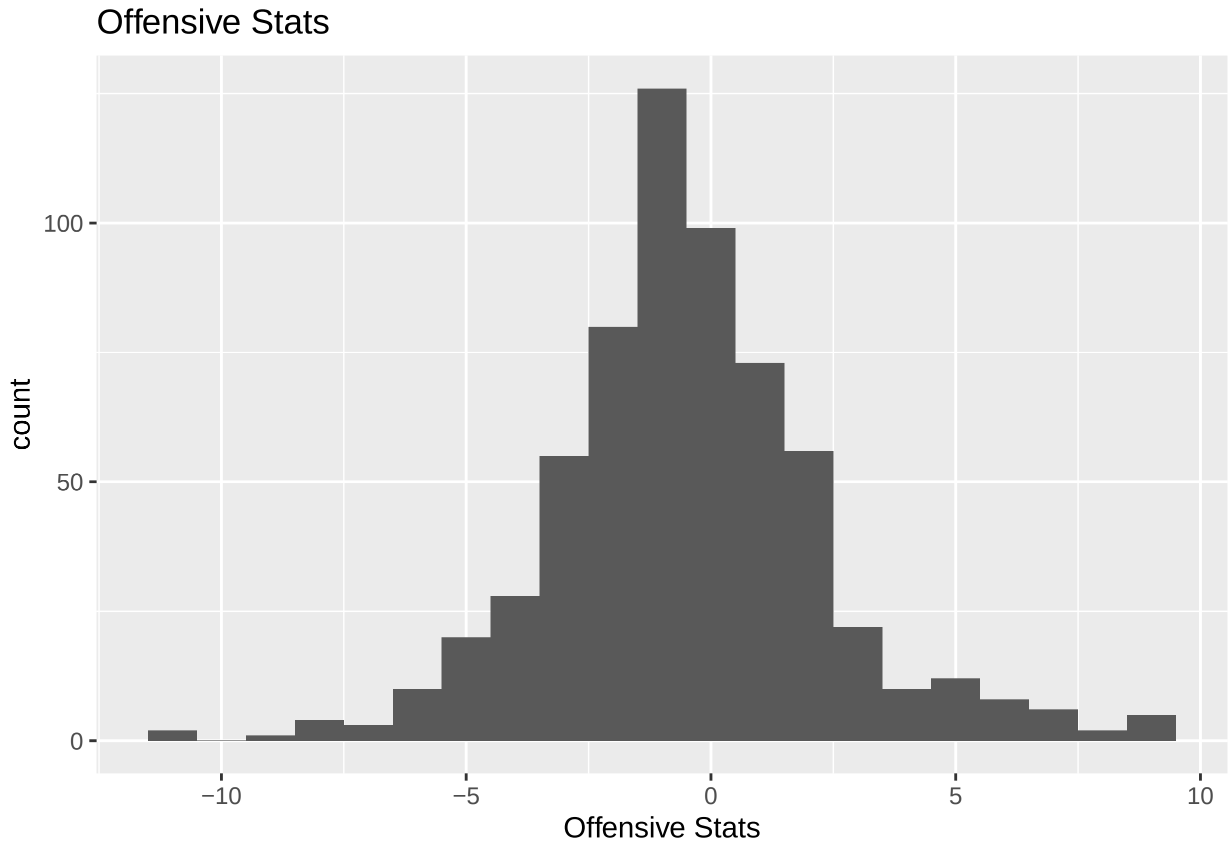
$$\widehat{RaptorDefense} = -0.33 + 0.00023\ MinutesPlayed$$

Intercept: A player without a single minute of playing time, would have a defensive score of -0.33. This is meaningless in context.

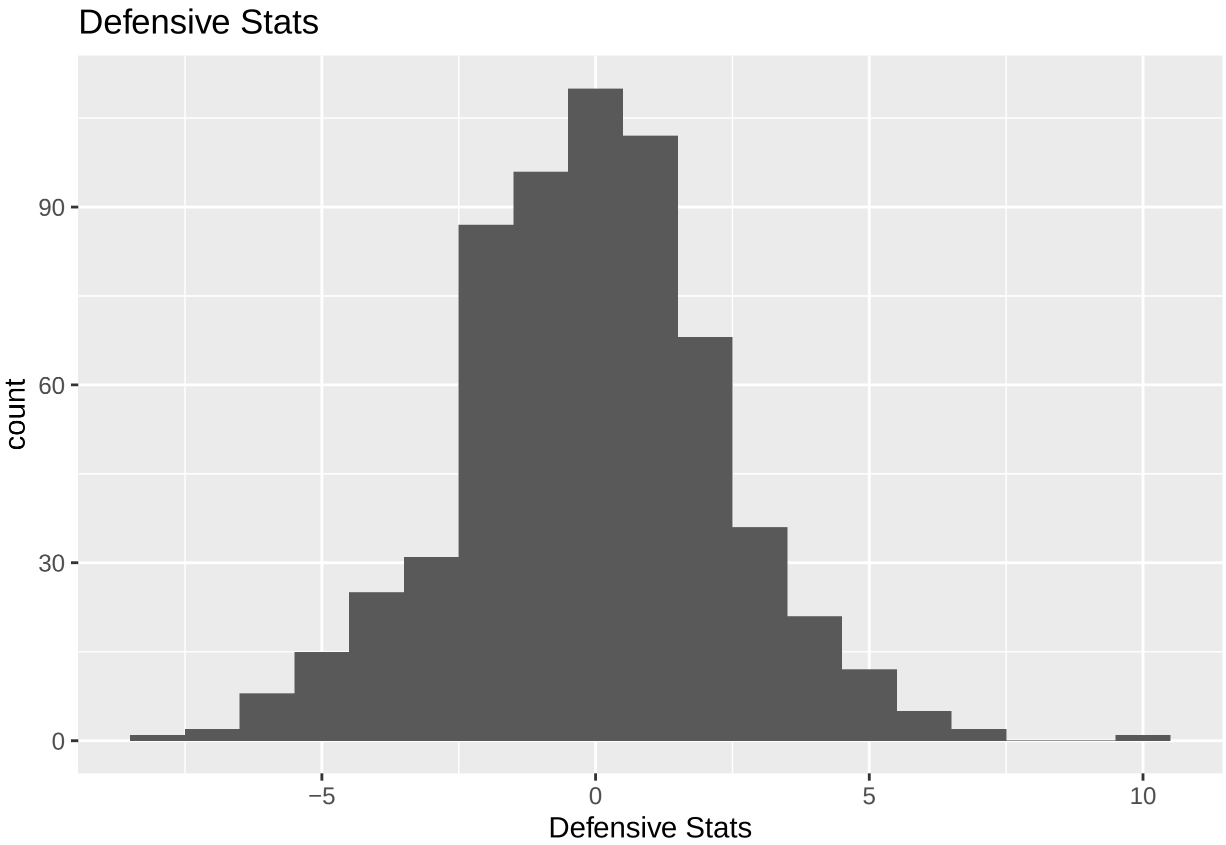
Slope: A one minute increase of playing time would predict a 0.00023 percent defensive score increase.

This data indicates that there is a slight correlation between playing time and good defensive stats.

The adjusted r-squared value of 0.00413 is low and indicates weak effect size.



This histogram reveals a normal distribution of offensive stats.



This histogram reveals a normal distribution of defensive stats.

Part 2

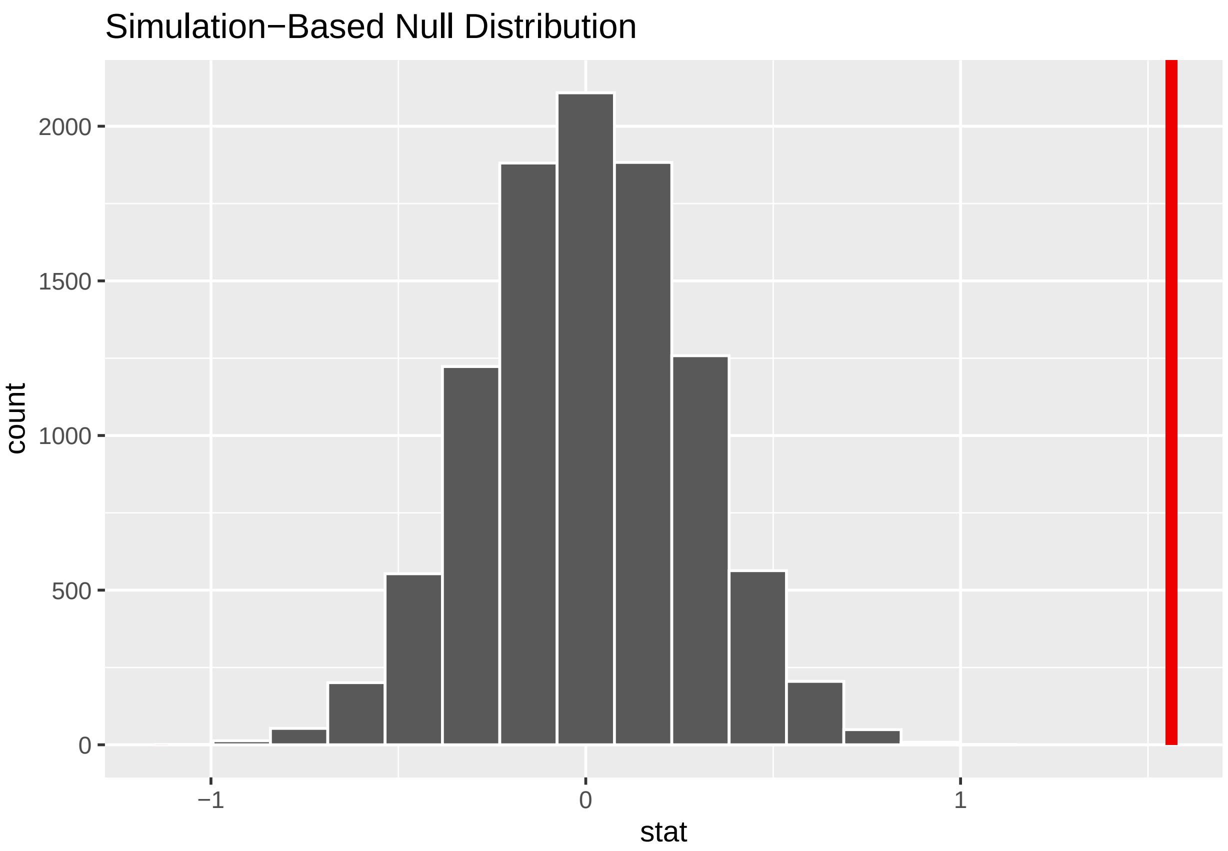
Are there discrepancies between the mean offensive and defensive scores based on if they were taken from the regular season or the playoffs?

Based on the summary statistics above, the null hypothesis is that there is no difference between the mean offensive statistics in the regular season versus the playoffs. The alternative hypothesis is that there is a significant difference between the offensive statistics in the regular season versus the playoffs.

$H_0 : \mu_O r - \mu_O p = 0$ $H_A : \mu_O r - \mu_O p \neq 0$

```
## # A tibble: 2 x 2
##   season_type   mean
##   <chr>        <dbl>
## 1 PO           0.767
## 2 RS          -0.796
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```



The p value here is 0.00, which when using an alpha of 0.05, reveals that we do reject the null hypothesis that there is no difference in the offensive statistics between the regular season and playoffs. This means that there is a statistically significant difference between the average offensive statistics between the regular season and the playoffs.

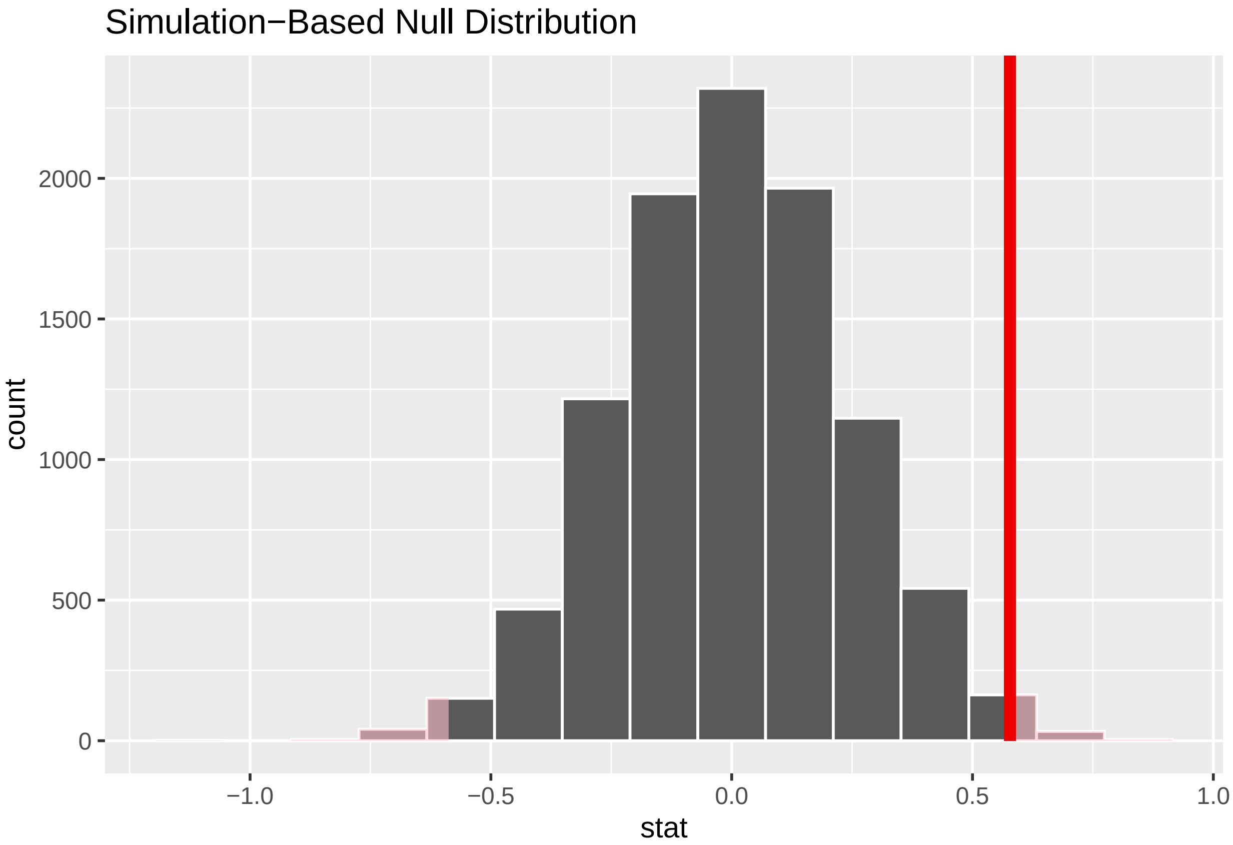
We will now investigate this same question but in regards to defensive statistics.

Based on the summary statistics above, the null hypothesis is that there is no difference between the mean defensive statistics in the regular season versus the playoffs. The alternative hypothesis is that there is a significant difference between the defensive statistics in the regular season versus the playoffs.

$H_0 : \mu_D r - \mu_D p = 0$ $H_A : \mu_D r - \mu_D p \neq 0$

```
## # A tibble: 2 x 2
##   season_type   mean
##   <chr>        <dbl>
## 1 PO           0.361
## 2 RS          -0.217
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0154
```



Here we have a p value of 0.0154, which is again smaller than the alpha of 0.05, so once again we reject the null hypothesis that there is no significant difference between the average defensive statistics between the regular season and the playoffs.

Results

We arrived at answers to our first research question using linear regression and plotting scatterplots to understand the correlations and how much minutes played influenced player statistics. Using linear regression to help predict and model how minutes played impact offensive and defensive statistics illustrated that a singular minute increase did not substantially increase defensive or offensive statistics. However, offensive statistics increased by 0.0013 for every minute while defensive statistics only increased by 0.00052 which is significantly less. Further, the adjusted R-squared values are incredibly low, indicating weak causation, so the linear regressions are not incredibly effective in context.

For the second question, we got a p-value close to 0 and 0.0154 for our hypothesis tests, indicating that we can safely reject the null hypothesis and further investigate the alternative hypothesis i.e., there is a difference between average defensive and offensive statistics for the regular season and the playoffs.

These methods were chosen to effectively analyze that the correlations between minutes played and increased statistics were moderately positively correlated, but there are other influences other than minutes played that impact offensive and defensive statistics, as seen by R-squared values. The small R-squared values of 0.06 for offensive and 0.004 for defensive statistics illustrate that several other factors hold bigger influences on improving statistics and playing better than just minutes played in games in the regular season and post season.

Discussion

We explored the following research questions:

Are there meaningful correlations between minutes played and offensive or defensive raptor scores defensive scores during the 2018 season?

Further, we compared whether or not these conclusions are the same if the scores are taken from the regular season (RS) or the playoffs (PO).

To answer the first question, we graphed raptor score for both categories against minutes played, and we also ran a linear regression on minutes as an explanatory statistic for raptor score. We learned that an increase of 1 minute led to a 0.00023 and 0.0009 increase in raptor defensive and offensive score, respectively.

From these numbers we can see there is a slight correlation between minutes played and raptor scores, but it doesn't appear to be too significant.

To answer the second research question, we filtered our data set by playoff and regular season games to see if there was a significant difference between regular season and raptor playoff scores. We calculated the difference in means between the two categorical variables and ran an independence test. Our p-value was less than 0.05 for both offensive and defensive raptor scores. The value indicates a statistically significant conclusion can be drawn between raptor score and the categorical variable, season type.

Overall, we learned a lot. First, we learned that offensive and defensive statistics rely on various factors. In a sport like basketball where scoring also depends on the player's position (guards usually score higher than centres, for example), it is difficult to draw conclusions using just numerical data. We also learned that contrary to our initial hypothesis, offensive and defensive statistics do actually depend on post-season/regular season. This is unsurprising given that players tend to 'clutch up' or perform better in the post-season. However, since the data set is smaller (fewer teams playing), the results should be examined carefully.

One critique of our method could be a lack of data points. We are only looking at a single season's data, so we could use more than just one seasons data. For instance, the 2018 season might have been a weird playoff season where the players performed uncharacteristically well (which they probably did since Lebron James and the Cavaliers got swept in 4 games). The conclusions to the research questions could be drawn from a greater number of seasons, rather than just the 2018 data. Besides the limited data points, the validity of the results are strong. It's clear that minutes played has a slight impact on raptor score albeit a marginal one. Additionally, it is extremely clear that there is a statistically significant difference for raptor score based on the season type.

I think if we were to do anything differently it would be to add more tests or potentially explore more questions. The use of some sort of plotting technique to analyse the second research question could have also been helpful Those would be two areas that we could've improved in.