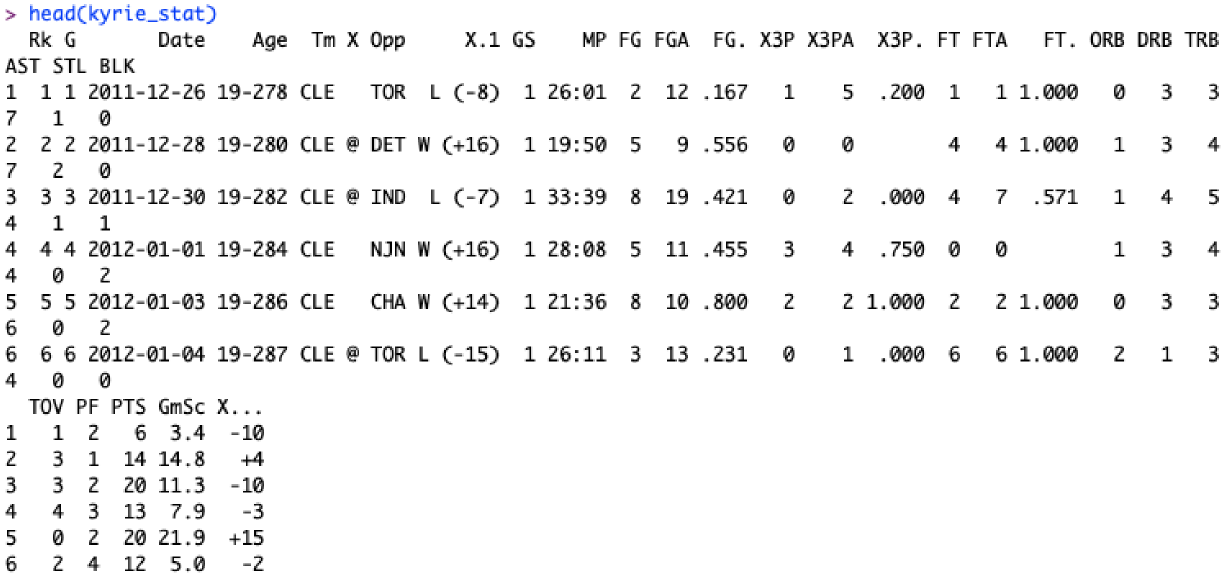
KYRIE IRIVING PERFORMANCE ANALYSIS

1. INTRODUCTION

- The goal of this project is to provide readers with insight about my favorite basketball player in the NBA - Kyrie Irving. Utilizing the data found on the internet, I try to gather insight of Kyrie's performance over time and what factors are impacting his overall performance from 2011 to 2017.

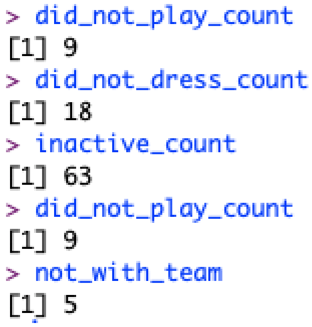
2. DATA UNDERSTANDING AND PREPROCESSING

- Kyrie was drafted in the NBA in 2011 and he has played in multiple teams since then. However, I decided that I will only analyze his statistics during the time he played for the Cleveland Cavaliers (2011 - 2017) which is also the longest time he played for a single team.
- In order to perform a Two Sample t-test, I also gathered data from his playoffs games starting from 2014. For context, the NBA has a thing called a regular season where each team plays against each other for a total of 82 games. After that, they are ranked based on their wins and losses and the team that has a good enough score will be chosen to play in the playoffs. Therefore, with this extra data, I can gain some insight about how Kyrie performs in different conditions (playoff games tend to be more tense and you also play against better teams).



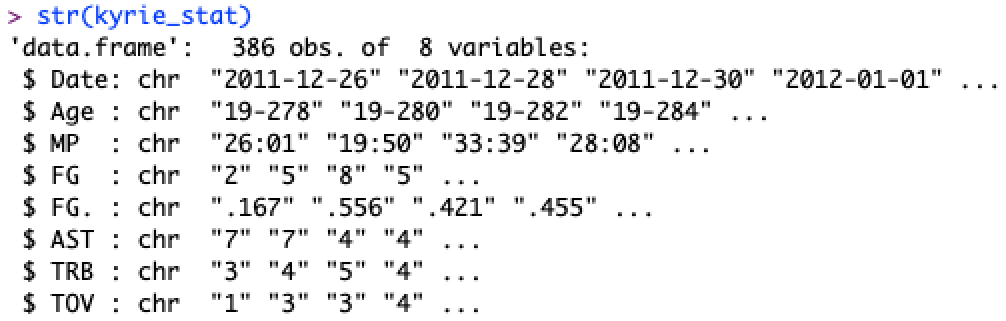
```
> head(kyrie_stat)
  Rk G       Date    Age  Tm X Opp     X.1 GS    MP FG FGA  FG. X3P X3PA  X3P. FT FTA   FT. ORB DRB TRB
AST STL BLK
1  1 1 2011-12-26 19-278 CLE   TOR  L (-8)  1 26:01  2  12 .167   1    5  .200  1   1 1.000   0   3   3
7   1   0
2  2 2 2011-12-28 19-280 CLE @ DET W (+16)  1 19:50  5   9 .556   0    0        4   4 1.000   1   3   4
7   2   0
3  3 3 2011-12-30 19-282 CLE @ IND  L (-7)  1 33:39  8  19 .421   0    2  .000  4   7  .571   1   4   5
4   1   1
4  4 4 2012-01-01 19-284 CLE   NJN W (+16)  1 28:08  5  11 .455   3    4  .750  0   0         1   3   4
4   0   2
5  5 5 2012-01-03 19-286 CLE   CHA W (+14)  1 21:36  8  10 .800   2    2 1.000  2   2 1.000   0   3   3
6   0   2
6  6 6 2012-01-04 19-287 CLE @ TOR L (-15)  1 26:11  3  13 .231   0    1  .000  6   6 1.000   2   1   3
4   0   0
  TOV PF PTS GmSc X...
1   1  2   6  3.4  -10
2   3  1  14 14.8   +4
3   3  2  20 11.3  -10
4   4  3  13  7.9   -3
5   0  2  20 21.9  +15
6   2  4  12  5.0   -2
```

- We can see that there are a lot of columns in the dataset so I decided to only use the most important ones: Age, Date, MP (minutes played), AST, TRB (total rebound), TOV (turnover - number of times he lost the ball), PTS



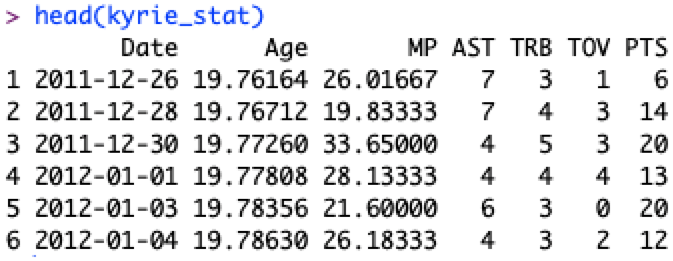
```
> did_not_play_count
[1] 9
> did_not_dress_count
[1] 18
> inactive_count
[1] 63
> did_not_play_count
[1] 9
> not_with_team
[1] 5
```

- Then after checking the data, we can see that there are some games that Kyrie didn't play because of suspension, injuries, … so I have to remove those entries from the data set.

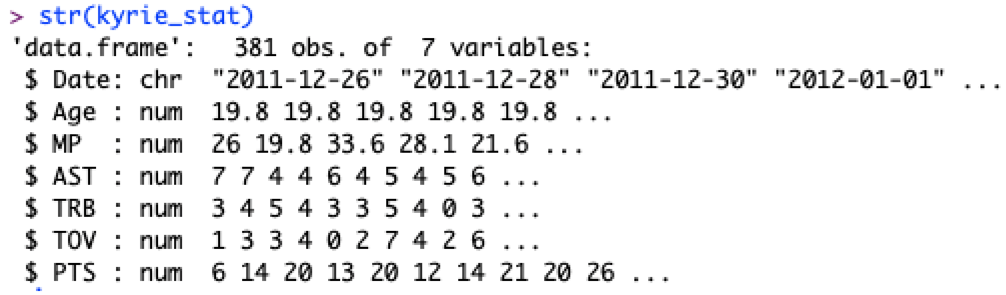


```
> str(kyrie_stat)
'data.frame':	386 obs. of  8 variables:
 $ Date: chr  "2011-12-26" "2011-12-28" "2011-12-30" "2012-01-01" ...
 $ Age : chr  "19-278" "19-280" "19-282" "19-284" ...
 $ MP  : chr  "26:01" "19:50" "33:39" "28:08" ...
 $ FG  : chr  "2" "5" "8" "5" ...
 $ FG. : chr  ".167" ".556" ".421" ".455" ...
 $ AST : chr  "7" "7" "4" "4" ...
 $ TRB : chr  "3" "4" "5" "4" ...
 $ TOV : chr  "1" "3" "3" "4" ...
```

- We can see that the data at first aren't recognized by R as numbers so I have to manually convert each column. Age and MP are special as they are in a unique format so I have to write different functions in order to extract meaningful data from them.

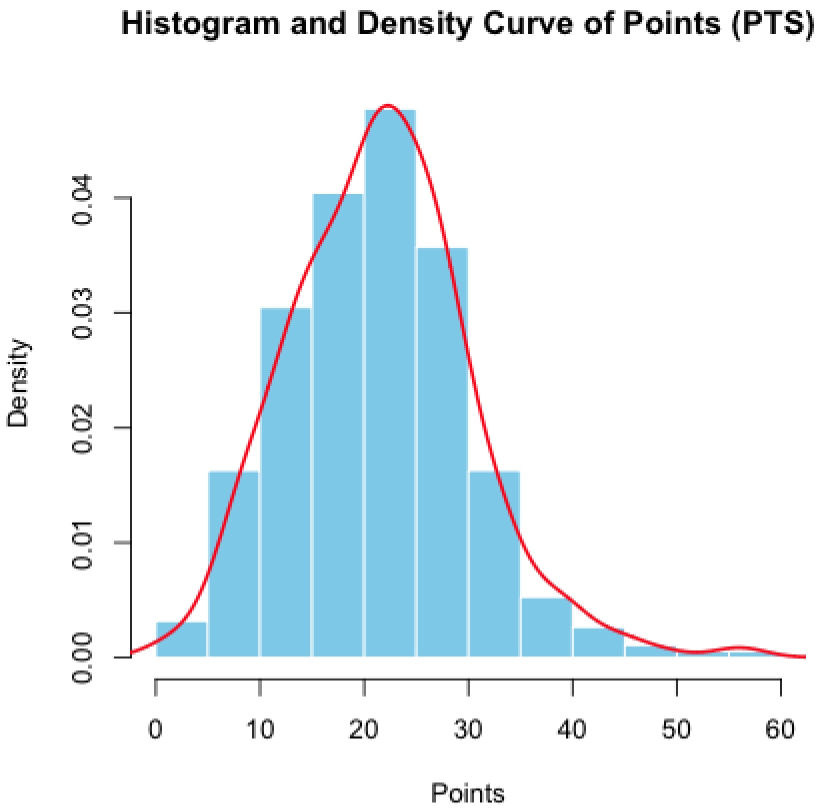


```
> head(kyrie_stat)
        Date      Age       MP AST TRB TOV PTS
1 2011-12-26 19.76164 26.01667   7   3   1   6
2 2011-12-28 19.76712 19.83333   7   4   3  14
3 2011-12-30 19.77260 33.65000   4   5   3  20
4 2012-01-01 19.77808 28.13333   4   4   4  13
5 2012-01-03 19.78356 21.60000   6   3   0  20
6 2012-01-04 19.78630 26.18333   4   3   2  12
```

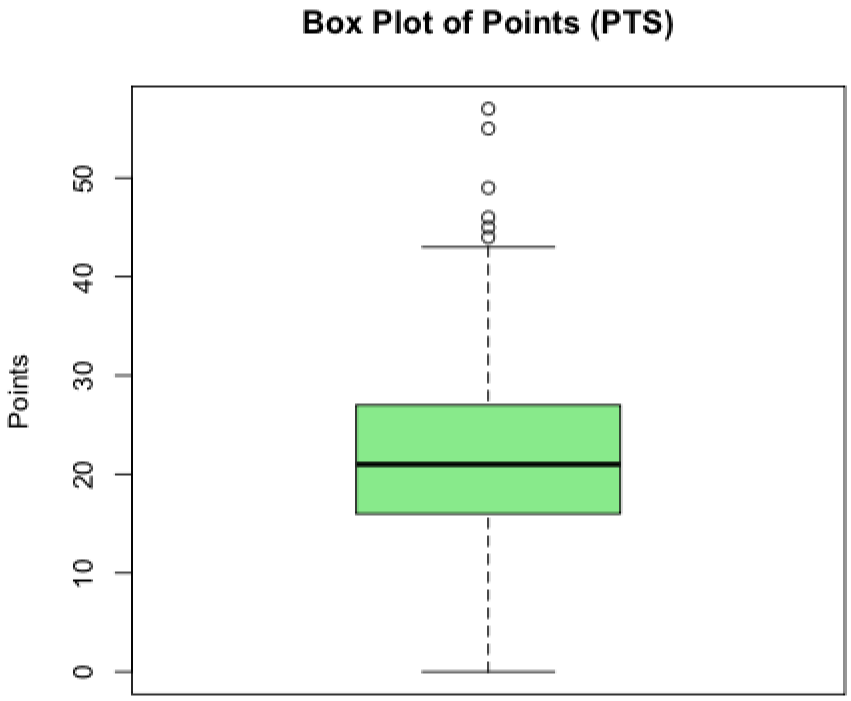


```
> str(kyrie_stat)
'data.frame':	381 obs. of  7 variables:
 $ Date: chr  "2011-12-26" "2011-12-28" "2011-12-30" "2012-01-01" ...
 $ Age : num  19.8 19.8 19.8 19.8 19.8 ...
 $ MP  : num  26 19.8 33.6 28.1 21.6 ...
 $ AST : num  7 7 4 4 6 4 5 4 5 6 ...
 $ TRB : num  3 4 5 4 3 3 5 4 0 3 ...
 $ TOV : num  1 3 3 4 0 2 7 4 2 6 ...
 $ PTS : num  6 14 20 13 20 12 14 21 20 26 ...
```

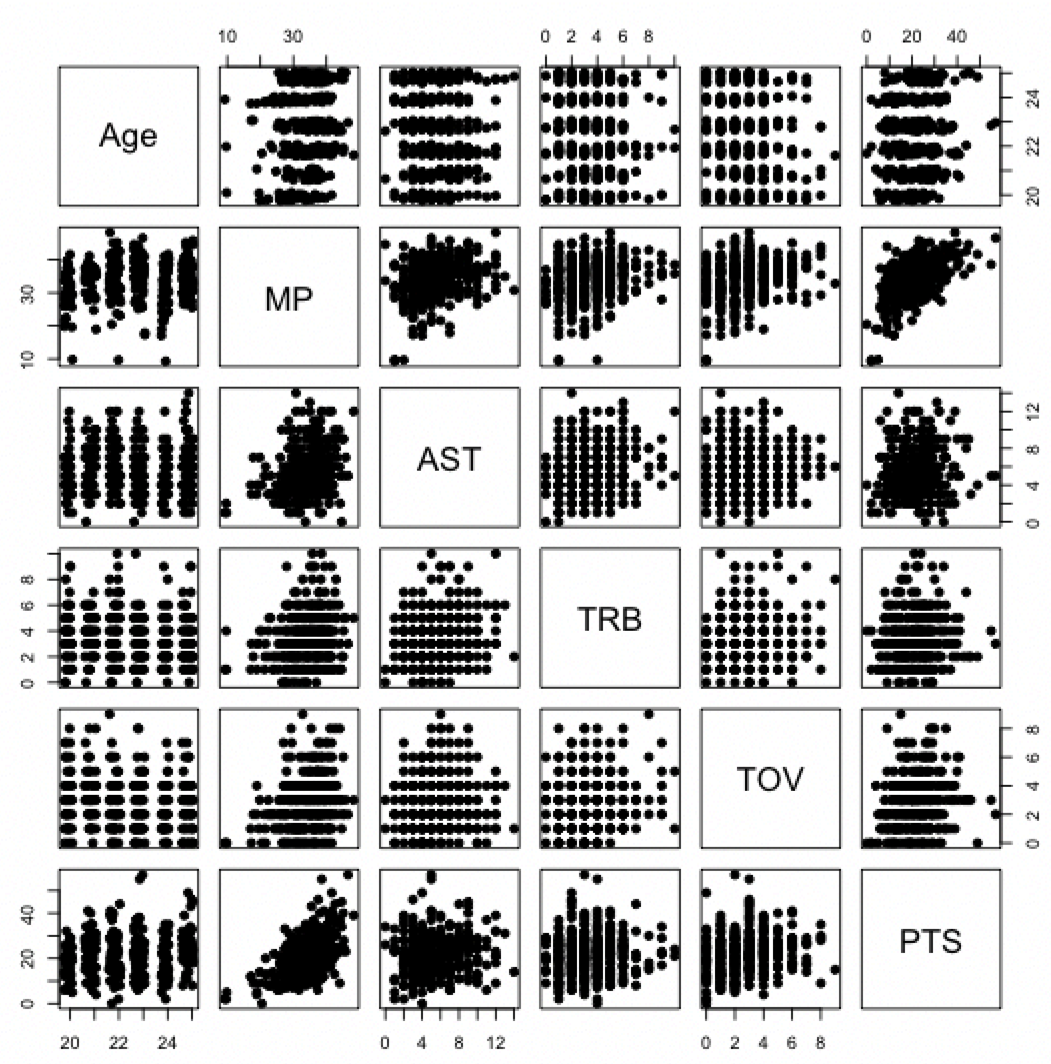
3. DATA OVERVIEW



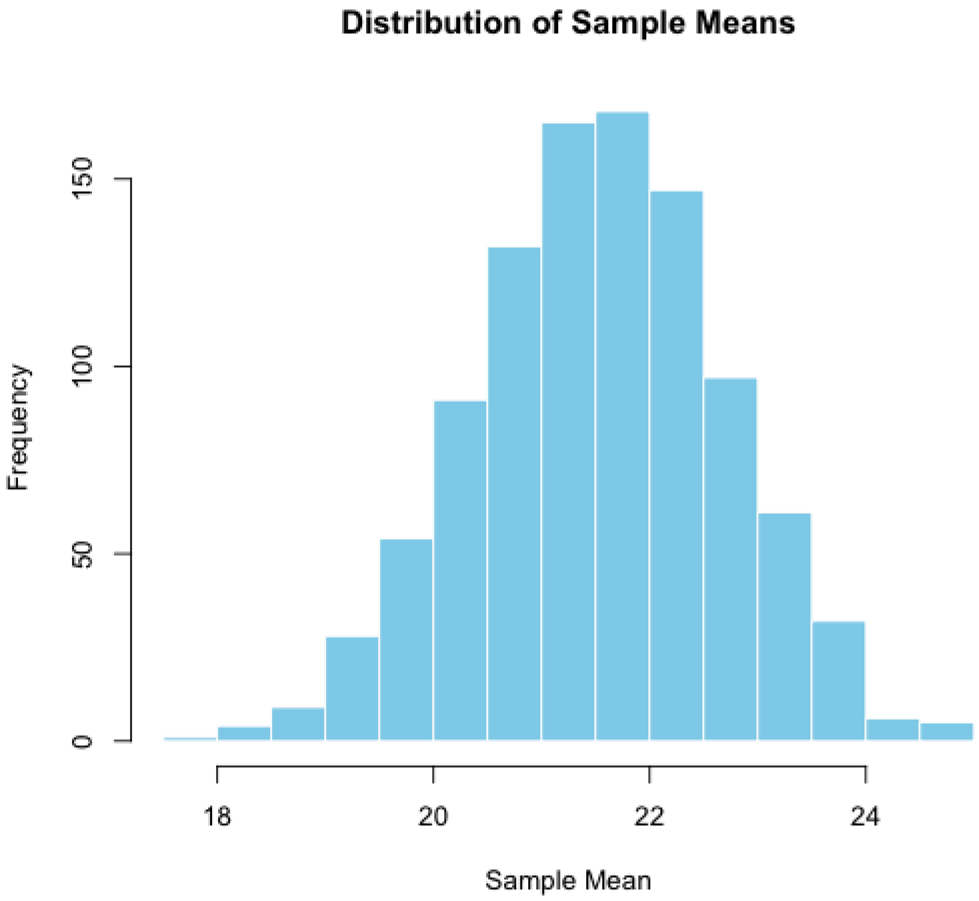
- From the histogram, we can see that the data is right-skewed so the mean is expected to be greater than the median. We can also see that the density is really low at the start of the points and at the end of the points at around 45 to 60 points. We can see that Kyrie is an excellent player as his scores are always around 20 to 30 points which is normally around 25-30% of the entire team's scores.

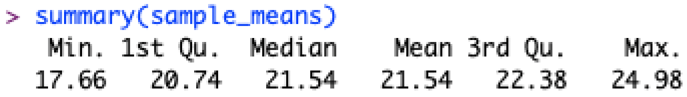


- From the boxplot, we can see that Kyrie’s points mostly range from around 16 points to 28 points. We can also see 6 outliers according to R with the games in which he scored over 45 points (which in basketball often referred as Career High).



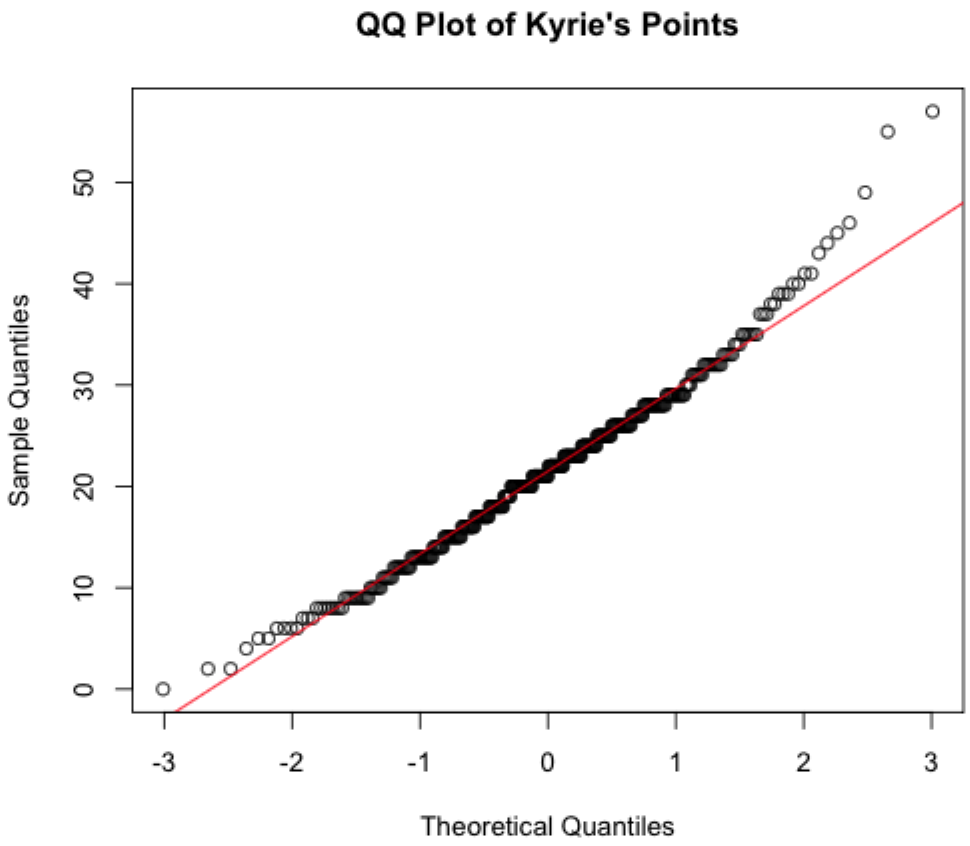
- Unfortunately, the scatter plots used to find the relationship of the pair of variables do not give a good result which is somewhat expected as basketball is a dynamic sport.
- We can still see a positive relationship between MP and PTS, MP and AST, .. which are somewhat expected as you have more time to get more scores, get more assists, … during the game.
- MP and PTS show the strongest relationship out of all the pairs.
- Age and PTS doesn’t make much sense here because age is not an integer otherwise it should also give a strong relationship





```
> summary(sample_means)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17.66   20.74   21.54   21.54   22.38   24.98 
```

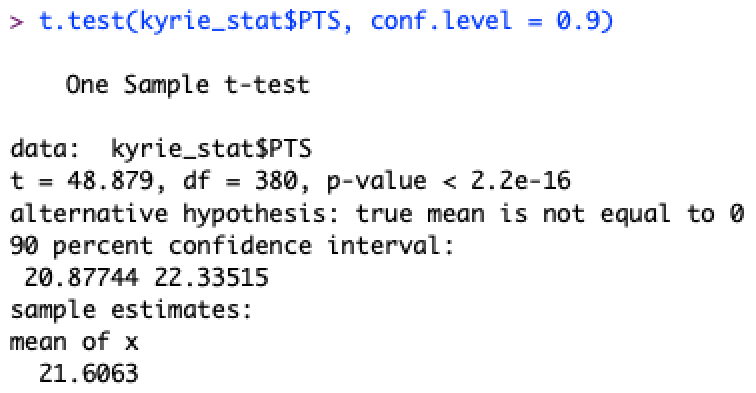
- I also try to use mean sampling with 1000 samples and sample size of 50 to see if Kyrie’s performance is consistent if we only look at 50 games (a lower sample size will give a more dynamic result as Kyrie had played 381 games during that time)
- As you can see, the histogram has a bell curve shape which means that Kyrie scored consistently in most of his games.



- Plotting Kyrie’s points also suggests that it is somewhat a normal distribution and is reliable to use the Central Limit Theorem although there are some fluctuations because of his Career High games and because he also played better after being in the NBA for longer.

4. IN-DEPTH ANALYSIS

- Now I want to see the confidence interval of Kyrie’s score to see the chances of him playing well.



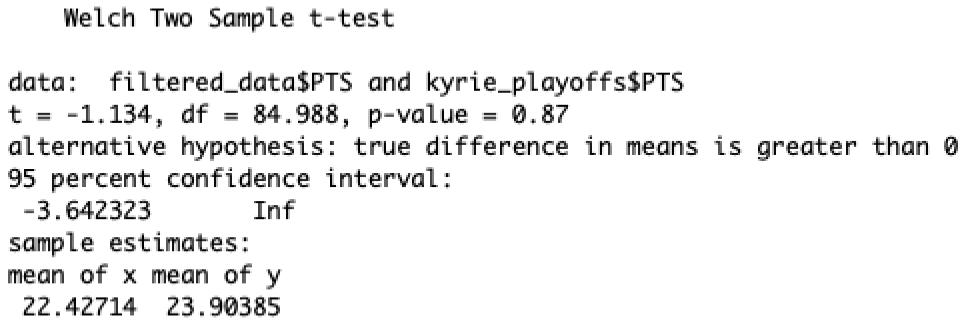
```
> t.test(kyrie_stat$PTS, conf.level = 0.9)

	One Sample t-test

data:  kyrie_stat$PTS
t = 48.879, df = 380, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 20.87744 22.33515
sample estimates:
mean of x 
  21.6063 
```

- We can see that he has 90% chances to score from 20.8 points to 22.3 points each game which should put him in the top in the NBA (he is in fact has been in the top 30 scoring leaders in the NBA for at least 6 times using available data up to now). This showcases his potential even in his early career.

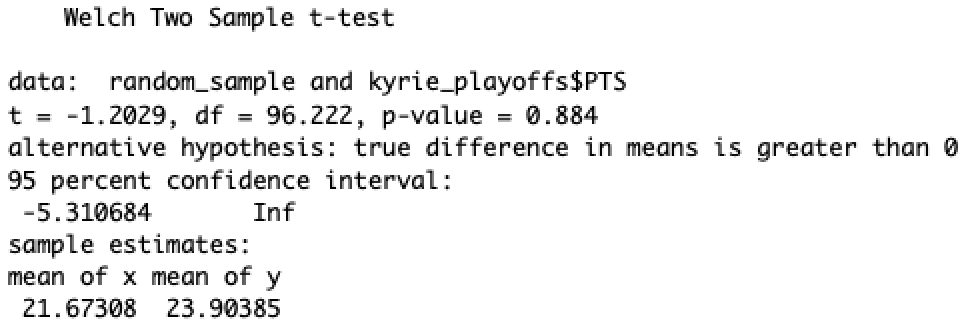
- I also want to see how his scores are affected using the playoff dataset I have gathered before. I want to know if he can still perform well when playing under higher pressure and against better teams.
- Because he still played nearly 4 times more during the regular season in comparison with the playoffs, 199 games and 52 games respectively. So I decided to try using 2 sample t-test with a random sample of size 52 of the regular season data and one with the whole regular season data.



```
	Welch Two Sample t-test

data:  filtered_data$PTS and kyrie_playoffs$PTS
t = -1.134, df = 84.988, p-value = 0.87
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -3.642323       Inf
sample estimates:
mean of x mean of y 
 22.42714  23.90385 
```

- According to the test, we can see that p-value = 0.87 which is a lot greater than the alpha value of 0.05 so we fail to reject the null hypothesis. Therefore, there is no sufficient evidence that Kyrie performed better in regular seasons. This is actually a really cool insight for me that everyone without using statistics said that he played worse in the playoffs.

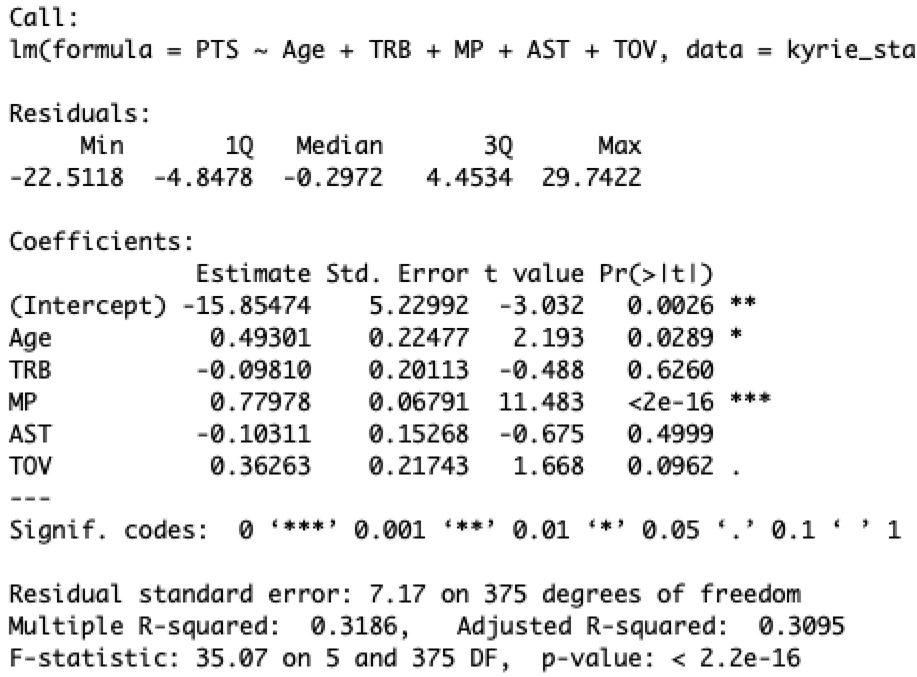


```
	Welch Two Sample t-test

data:  random_sample and kyrie_playoffs$PTS
t = -1.2029, df = 96.222, p-value = 0.884
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -5.310684       Inf
sample estimates:
mean of x mean of y 
 21.67308  23.90385 
```

- Even with a random sample, there is no evidence that he played worse during the playoffs.

- For my final question, I want to know which attribute of his games that has an influence on his scoring performance using Linear Regression and Backstepping



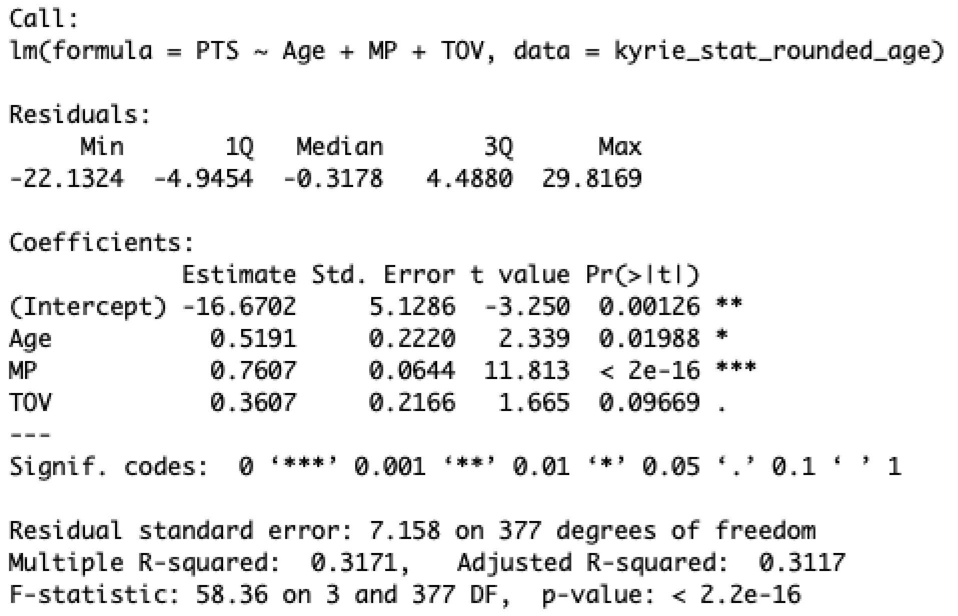
```
Call:
lm(formula = PTS ~ Age + TRB + MP + AST + TOV, data = kyrie_stat_rounded_age)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.5118  -4.8478  -0.2972   4.4534  29.7422 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -15.85474    5.22992  -3.032   0.0026 ** 
Age           0.49301    0.22477   2.193   0.0289 *  
TRB          -0.09810    0.20113  -0.488   0.6260    
MP            0.77978    0.06791  11.483   <2e-16 ***
AST          -0.10311    0.15268  -0.675   0.4999    
TOV           0.36263    0.21743   1.668   0.0962 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.17 on 375 degrees of freedom
Multiple R-squared:  0.3186,	Adjusted R-squared:  0.3095 
F-statistic: 35.07 on 5 and 375 DF,  p-value: < 2.2e-16
```

- According to the summary, we can see that Age and MP have the most positive influence on Kyrie performance, which is natural as people played better the more they played and adapted to the NBA environment.
- We can also see that AST and TRB have a negative effect on his performance although not significant. This is normal as AST and TRB are the statistics related to helping the team.
- The R-squared value is 0.3186 and the Adjusted R-squared is 0.3095 which means that around 31% of the variance in Kyrie’s scores can be explained by the other attributes. I think this is normal because a player's performance is also greatly influenced by the events that happened in their life, injuries and illness which are not reflected inside the dataset.
- One weird thing in this model is that TOV has a positive relationship with PTS and is pretty close to being significant (0.0962). This is weird because TOV means that Kyrie lost the balls when playing which should normally negatively affect his performance.



```
Call:
lm(formula = PTS ~ Age + MP + TOV, data = kyrie_stat_rounded_age)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.1324  -4.9454  -0.3178   4.4880  29.8169 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -16.6702     5.1286  -3.250  0.00126 ** 
Age           0.5191     0.2220   2.339  0.01988 *  
MP            0.7607     0.0644  11.813  < 2e-16 ***
TOV           0.3607     0.2166   1.665  0.09669 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.158 on 377 degrees of freedom
Multiple R-squared:  0.3171,	Adjusted R-squared:  0.3117 
F-statistic: 58.36 on 3 and 377 DF,  p-value: < 2.2e-16
```

- After performing backstepping, only Age, MP and TOV are left in the equation which is somewhat expected
- We can see that the R-Squared value doesn't change much (0.3171 compared to the original 0.3186).