# Statistical analysis on the NBA(2014-15) dataset

Siva Sathguru Pandiyarajan

12 August 2021

**Abstract**

The aim of this report is on showing how different descriptive and inferential statistical approaches may be used to and how their findings can be interpreted and presented by using National Basketball Association (NBA), 2014-2015 dataset.The evaluation is established on the effect of accuracy over fatigue and stress along with the defender proximity analysis and shooting distance over shot accuracy,where we have analysed that the elite basketball players cope really well with such extreme circumstances however effects of psychological stress are highly visible. The goal of this research is to determine if the top scorer of a team is able to assist his side win the match, when both the number of shooting attempts and the accuracy are more exceptional. In addition, we analyse this season's finest defenders.

## Introduction

The objective of this report is to analyse and to provide the relevant results for the data selected for use in statistical and math analyses. These analyses have been done using RStudio tools. This study will analyse different areas, show how the analytics were performed, discuss the theory employed, show and evaluate results achieved. If applicable, the study will also suggest future research topics.

```
noquote('Variables names:')
```

```
## [1] Variables names:
```

```
print(names(nbaleague_df))
```

```
##  [1] "GAME_ID"            "DATE"               "HOME_TEAM"
##  [4] "AWAY_TEAM"          "PLAYER_NAME"        "PLAYER_ID"
##  [7] "LOCATION"           "WIN_LOSE"           "SHOT_NUMBER"
## [10] "PERIOD"             "SEC_REMAIN"         "SHOT_CLOCK"
## [13] "DRIBBLES"           "TOUCH_TIME"         "SHOT_DIST"
## [16] "PTS_TYPE"           "CLOSEST_DEFENDER"   "CLOSEST_DEFENDER_ID"
## [19] "CLOSE_DEF_DIST"     "SUCCESS"
```

**Remark** The dataset consists of the home teams and away teams with the total of 902 games played over 120 days with 468 players appears as defenders with only 281 players appear as shooters across 30 different teams.

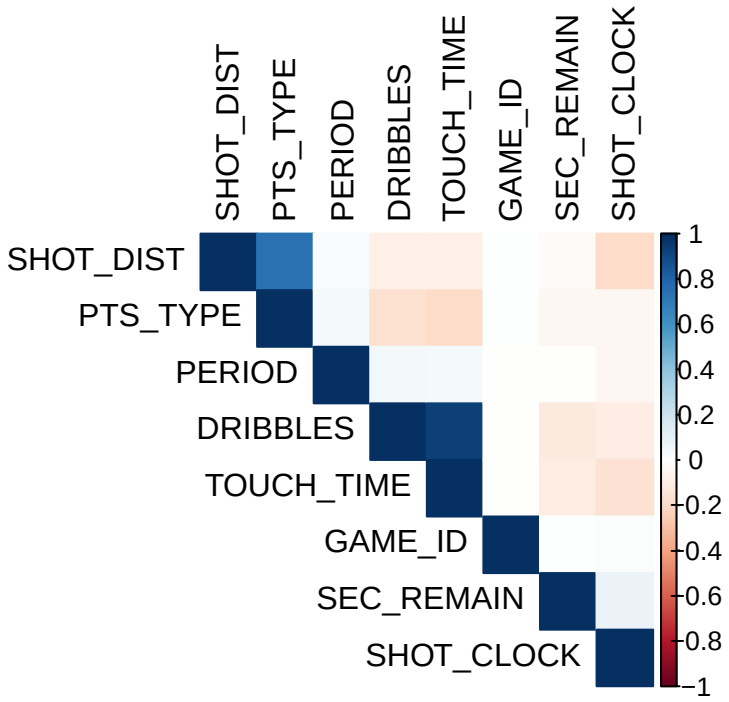
## Statistical Analysis

Descriptive analysis is to describe basic details of dataset. The quantitative description of number values for example median may be described as centrality measures, where the measurement of variation is spoken of standard deviation, variance, quartile ranges. We may utilize graphic plots to focus on the results of such a study. By summarizing it, it simplifies vast amounts of information sensibly. When we locate a mean,median and standard deviation, we can just tell much about a dataset.

Table 1: Summary of important numerical variables

| COL_NAME | MEAN_VALUE | MEDIAN_VALUE | S.D | I_Q_R |
|---|---|---|---|---|
| SHOT_NUMBER | 6.47962 | 5.0 | 4.640472 | 6.0 |
| SHOT_CLOCK | 12.21167 | 12.1 | 5.895360 | 8.6 |
| DRIBBLES | 2.04920 | 1.0 | 3.488069 | 3.0 |
| SHOT_DISTANCE | 13.49515 | 13.4 | 8.871681 | 17.8 |
| CLOSE_DEF_DIST | 4.11256 | 3.7 | 2.728465 | 3.0 |

Corrplot is a graphical representation of correlation matrix. Helps us to determine the cor-related attributes. We have utilized all the numerical values assigned to the graph but there are certain variables that are both category and numerical, as indicated, so that they could be connected or not.




As seen in the figure the Dribbles and touch_time has positive correlation as it shows a very favorable link between the amount of dribbles before shoot and touch time. This is quite acceptable, given the values basically measure the ball

before a player shoots. This shows if the player was dribbling before shooting. This variable has more predictive ability than any dribble or contact time and resolves the colinearity problem between both.

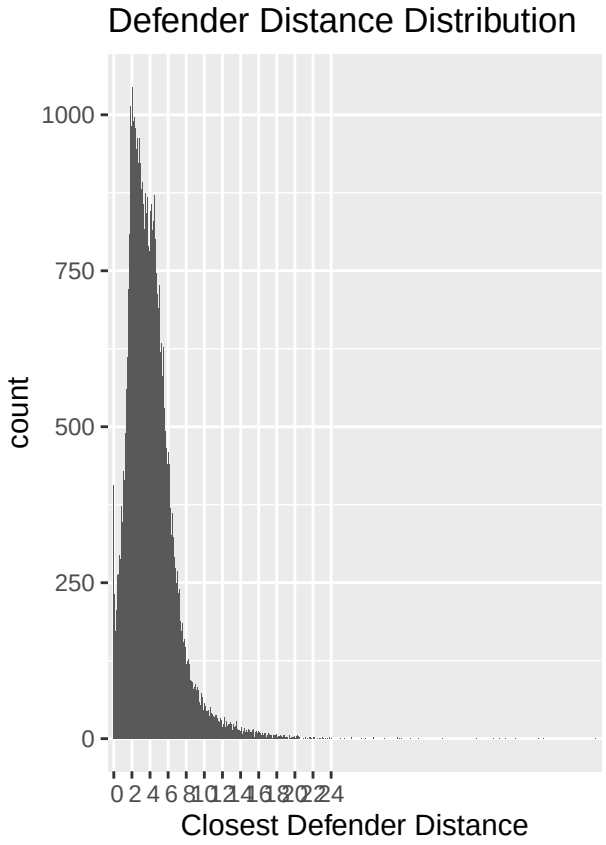
## Effects on Accuracy

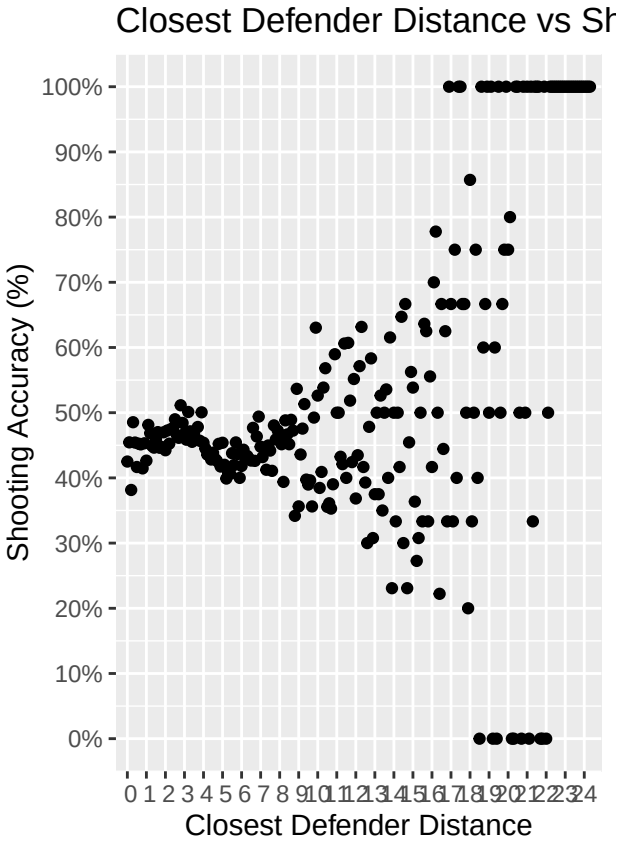
In this section, we investigate variables affect shot accuracy and shot selection.For each observation, approximately 20 different variables were tracked, including shooters distance to the basket, distance from the shooter to the nearest defender, and much more.

### Closest Defender Distance vs Shot Accuracy

This analysis focuses on whether there is a significant effect from closest defender distance to the shot outcome. The analysis is aimed to establish whether being away from the defender, i.e. having more time to prepare and execute a shot enables the shooter to score with a higher probability. Opposite effect will be when defender is close to the shooter it doesn't leave the shooter enough time to prepare and execute the shot properly.

The Null Hypothesis is that the distance to a closest defender does not have an impact on accuracy of shooting.




**Results**

Plot Result

**Table 1**: Relationship between the defender distance and accuracy

| Defender Distance (range in feet) | accuracy (range in %) |
|---|---|
| 0 - 2 | 38 - 49 |
| 2 - 4 | 44 - 50 |
| 4 - 6 | 41 - 46 |
| 6 - 8 | 39 - 50 |
| 8 - 10 | 34 - 54 |
| 10 - 16 | 24 - 66 |
| 16+ | 0 - 100 |

**Correlation Test**

The next analysis calculates the correlation between both variables. If the analysis included all data points for every distance, the correlation result would be as follows: There is a medium to strong positive correlation between the defender distance and shooting accuracy. The Pearson's correlation coefficient for the relationship between the variables "Defender Distance" and "Accuracy" was r(242) = .46, p < .01, with a level of significance of p = .05.

**Correlation Test Result**

When analyzing the relationship between distance and the accuracy for the range of 0 to 8 feet distance the result of the correlation is different:
There was weak negative correlation between the defender distance and shooting accuracy in this range. The Pearson's correlation coefficient for the relationship between the variables Defender Distance and Accuracy was r(297) = -0.18, p < .01, with the level of significance of p = .05.

**Discussion**

The research demonstrates that the defender distance in specific ranges has been impacted by the accuracy rate. There is an adverse connection between distance and accuracy with the defender's distance from 0 to 8 feet, implying that players are more likely to miss when the defender comes closer. Data vary from 0 percent for the distance longer than 16 feet (not shot from that distance) up to 100 percent (assured successful shot from that distance). This implies that the findings of this analysis are insignificant and should not be taken into account. It also shows that only a small amount of shots were made for higher defender distances. Therefore, it would affect the accuracy calculation result. However, from the correlation perspective, the result would be different if the analysis included all of the data points versus only short distance data points. This was arguably due to the fact that the defender distance does not mean the shooter distance from the basket was closer. Therefore, we reject the Null Hypothesis, since the defender distance affected shooting accuracy.

## Shooting Distance vs Shot Accuracy

This analysis focuses on whether there is any significant effect resulting from shooting distance from basket on the shot outcome. The NULL hypothesis is that the distance to the basket does not influence outcome of a shot.

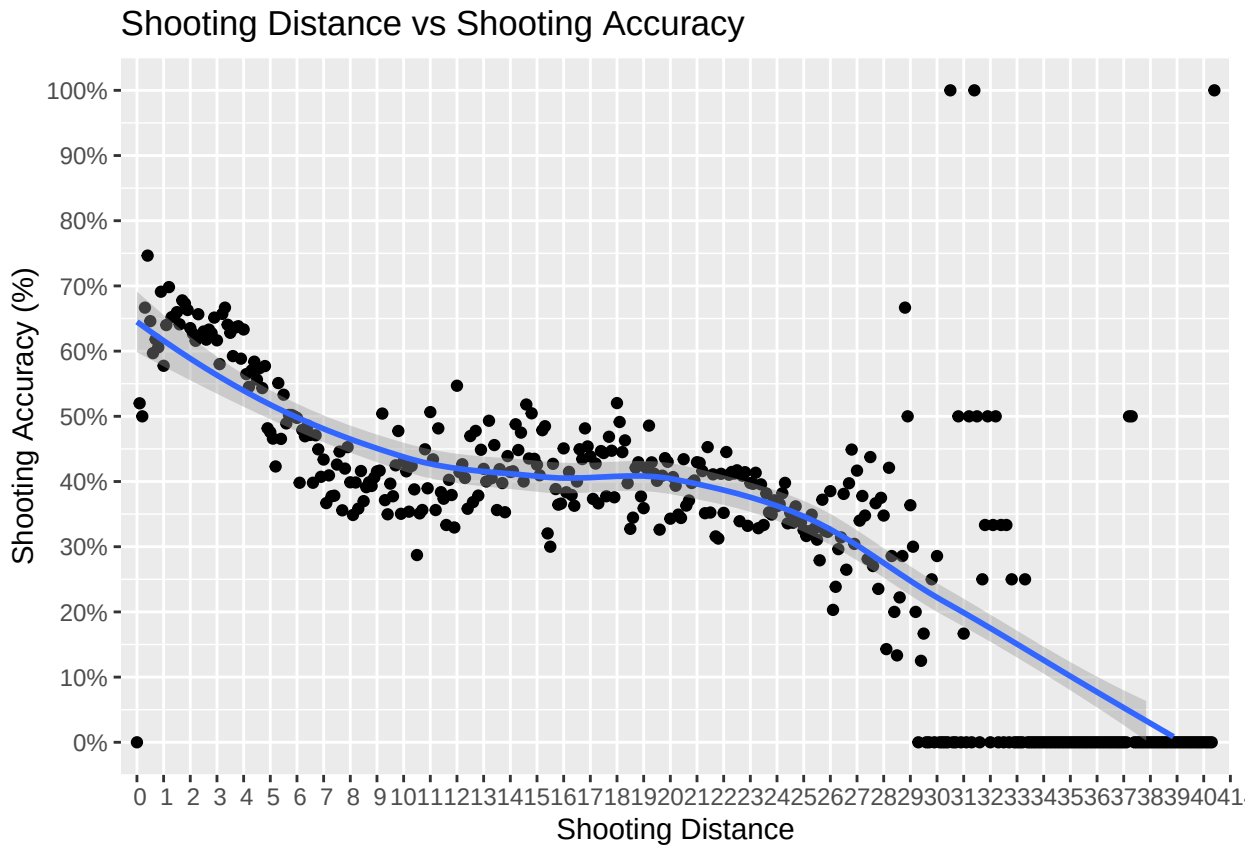



**Table 2**: The relationship between the shooter distance ranges and accuracy

| Distance (in feet) | Accuracy (range in %) |
|---|---|
| 0 - 2 | 50 - 70 |
| 2 - 4 | 60 - 66 |
| 4 - 6 | 47 - 57 |
| 6 - 8 | 35 - 49 |
| 8 - 10 | 35 - 51 |
| 10 - 12 | 28 - 45 |
| 12 - 14 | 33 - 55 |
| 14 - 16 | 35 - 52 |
| 16 - 18 | 37 - 52 |
| 18 - 24 | 31 - 51 |
| 24 - 28 | 20 - 45 |
| 28+ | 0 - 50 |

There is a strong negative correlation between the shooting distance and shooting accuracy. The Pearson's correlation coefficient for the relationship between the variables Shooting Distance and Shot Accuracy was r(403) = -24.41, $p < .05$, with the level of significance of $p = .05$.

**Correlation Test Result**

If the correlation test taken for a certain range of distance based on the diagram, the correlation

result would be different as follows: There was low positive correlation found between the shooting distance and shooting accuracy for the range 10 to 20 feet distance. The Pearson's correlation coefficient for the relationship between the variables Shooting Distance and Shot Accuracy was r(99) = 1.01, p =.31, failing to the level of significance (p = .05).

**Discussion**

Based on the schematics and the linkage above, the relationship between these factors is not exactly linear, but the firing distance influences the accuracy of the shoot. This was demonstrated in the chart in which, independent of the shooter skill, the accuracy for shots taken from 10 to 20 feet was comparatively similar. Regarding the poor success rate on longer distances, it might maybe be concluded that the shooter attempted to shoot at close-up time (buzzer beater) or almost zero in the shot clock or clock – the study would therefore reject the null hypothesis. The dataset mostly returns 0% for shooting distance greater than 30 feet. Therefore, the future research could be enriched with statistics from other seasons to see if these findings hold.

## Fatigue Effect (Quarter Accuracy Rate) & Shot Clock Pressure

This part of the analysis examines the potential impacts of fatigue and stress on the accuracy of shots.

Hypotheses for this area of analysis are: *Fatigue: The Null hypothesis states that there is no significant difference (p < .05) in the mean of shot accuracy in later periods of the game.* Stress: The Null hypothesis states that there is no significant difference (p < .05) in the means for the shot accuracy for lower seconds left on the shot clock.

**Remark** The T-tests for period 1 and period 2 was significant, t(25485) = 2.57, p = 0.009. Further the T-test for period 2 and 3 was not significant, t(25160) = -1.16, p = 0.244, as well as the T-test for period 1 and 3, t(26010) = 1.41, p = 0.15. The T-test for the first 3 quarter and last quarter was highly significant, t(24232) = 3.45, p < .01.

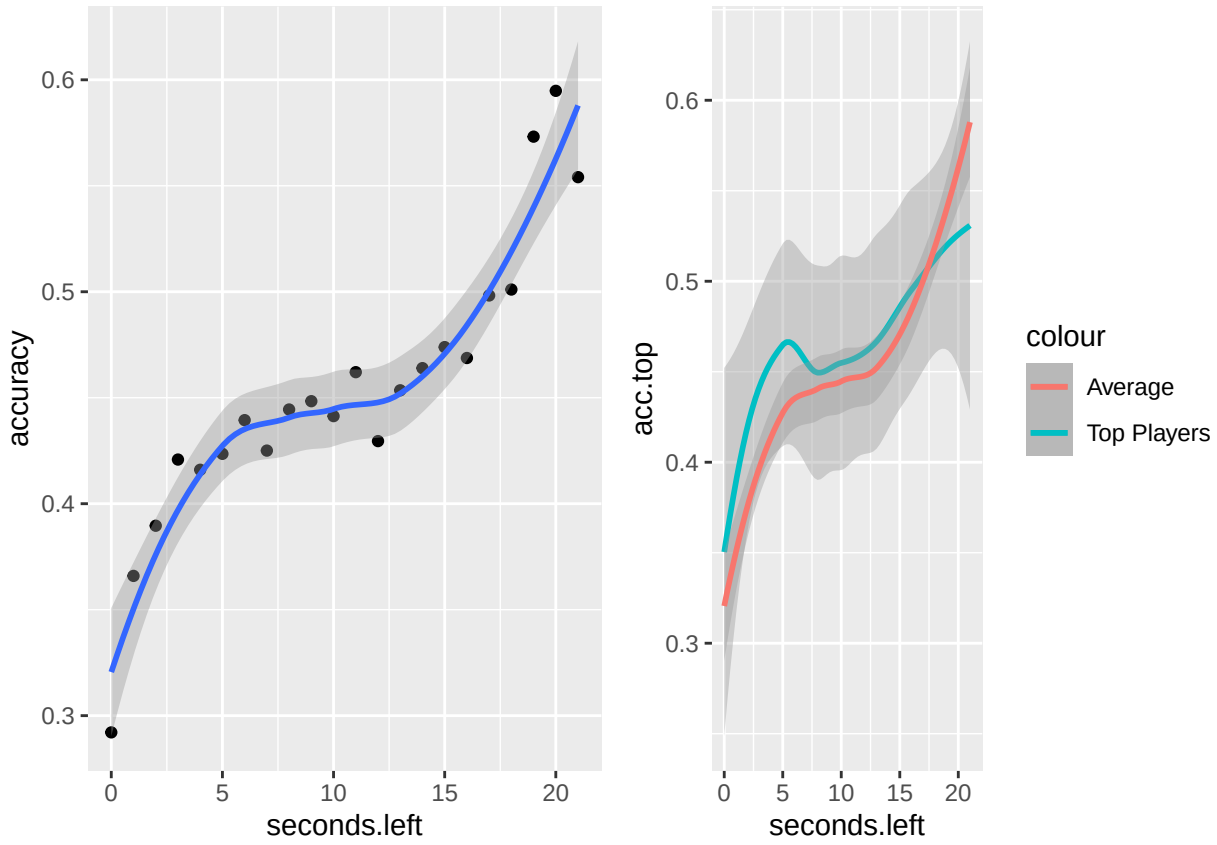
The changes in the overall accuracy per second are shown in the graph below.

After all, the Null hypothesis must be rejected as the statistical methods used in latter times of the game show differences of accuracy even if in the first quarter the game is small but accuracy for future periods cannot be concluded. A significant decline in precision over 3% (46 to under 43 percent) was observed in the fourth quarter, indicating players are tired and any overtime may be significantly harder.

The effects of psychological stress are, as shown in the chart above, extremely obvious. The accuracy of the shoot clock in the first ten seconds of the game time accuracy is higher than the normal accuracy (60%-45% vs. 45% in normal gambling), but it decreases quite quickly over the last five seconds of the shot clock (from 42% to 32%).

An ad-hoc data analysis was performed to see whether this pattern of stress-induced precision decrease can also be identified by the "super star" player. The two top 2 shooters of the season were James Harden and Kyrie Irving.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




The graph shows many similarities for top players to the average graph. The most visual differences are the slightly lower accuracy around the 20 seconds mark for the top players. Explanations for these differences could be that around the 20 second mark top players feel more obligated to shoot than the average player, even when they are not in the optimal position to do so - leading to lower accuracy.

## Game result impact by field goal attempt

**Introduction** The objective of this study is to see if a team's top scorer can aid the team to achieve the match by performing more effectively both in terms of shooting attempts and accuracy. The null hypothesis is that it is unlikely that a team's leading scorer will win with more field attempts. The second null hypothesis is that if a team's best scorer has a greater shooting % than average, it is unlikely that it would gain. As this is such a specialised study, a literature evaluation has not found enough data to cover this gap.

**Results**

1. Field Goal Attempts (FGA)

There was no correlation found between the outcome of the matches in this season and the number of attempts of the players made. The Pearson's correlation coefficient for the relationship between the variables WIN_LOSE and FGA was $r(841) < .01$, $p = .770$, failing to reach the level of significance ($p = .05$).

2. Field Goal Percentage (FG%)

There was no correlation found between the outcome of the matches in this season and the field goal percentage of the players. The Pearson's correlation coefficient for the relationship between the variables W and FGA was r(841) < .01, p = .110, failing to reach the level of significance (p = .05).

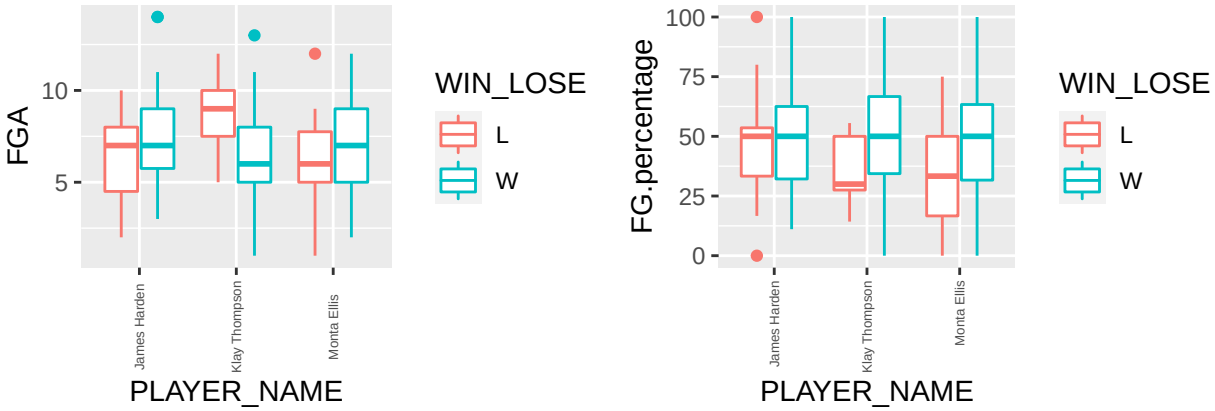


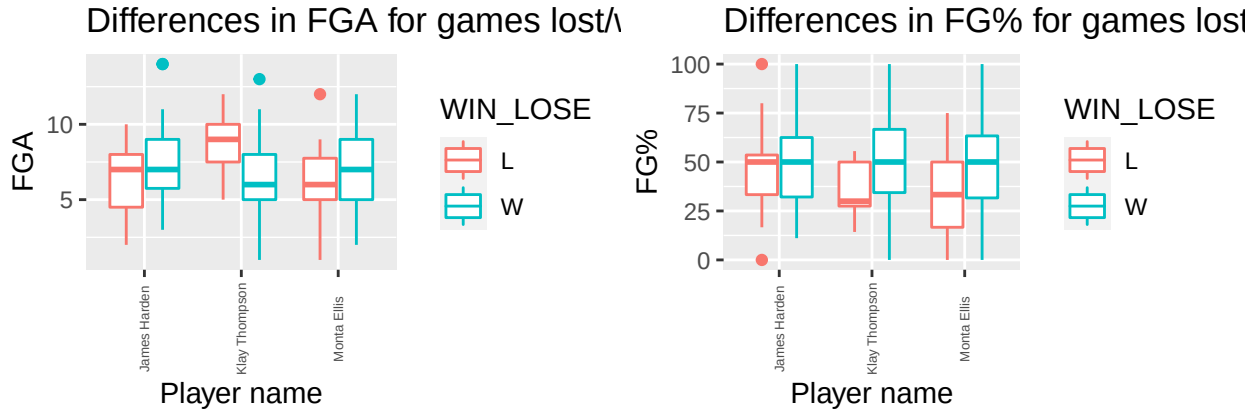

**Table 3**: Result for the logistic regression analysis for FGA and FG% of 15 players

| | Estimate | Std. Error | t value | Pr(>t) |
|---|---|---|---|---|
| (Intercept) | 0.5777166 | 0.0558336 | 10.347 | <2e-16 *** |
| FGA | -0.0027353 | 0.0062801 | -0.436 | 0.663 |
| FG.percentage | 0.0012903 | 0.0007914 | 1.630 | 0.103 |

**Table 4**: Result for the anova for FGA and FG% for 15 players

| | nbaleague_df | Deviance | Resid. nbaleague_df | Resid. Dev |
|---|---|---|---|---|
| NULL | | | 842 | 198.77 |
| FGA | 1 | 0.02009 | 841 | 198.75 |
| FG.percentage | 1 | 0.62692 | 840 | 198.12 |

An adhoc data analysis was used as 3 of the 15 players stood out in particular. To examine whether their data result in a better model, another logistic regression analysis was performed with the same dependent and independent variables for three special players: James Harden, Klay Thompson and Monta Ellis.




The box plots above shows the performances of the three special players in terms of FGA and FG% respectively.

**Table 5**: Result for the logistic regression analysis for FGA and FG% of the 3 special players

| | Estimate | Std. Error | t value | Pr(>t) |
|---|---|---|---|---|
| (Intercept) | 0.5244791 | 0.1162145 | 4.513 | 1.17e-05 *** |
| FGA | -0.0001002 | 0.0129524 | -0.008 | 0.99383 |
| FG.percentage | 0.0039459 | 0.0014703 | 2.684 | 0.00798 ** |

**Table 6**: Result for the anova for FGA and FG% for the 3 special players

| | nbaleague_df | Deviance | Resid. nbaleague_df | Resid. Dev |
|---|---|---|---|---|
| NULL | | | 176 | 36.723 |
| FGA | 1 | 0.00011 | 175 | 36.723 |
| FG.percentage | 1 | 1.45965 | 174 | 35.263 |

### Discussion

From the correlation tests between the match results and players' performances, it can be noticed that the increase in number of field goal attempts from top scorers is unlikely to make any impact on the results of the game. The p-values of both correlation tests are too high to reach the level of significance (p = .05).

From FGA and FG percent logistic regression analysis, it was revealed that FGA had no influence on the match result. FG% is an important indicator of the match result. The anova has shown, however, that both variables have a fairly significant residual deviation and hence additional study is required to enhance this model.

For these three players(James Harden, Klay Thompson and Monta Ellis), the p-values still did not reach the level of significance (p = .05) for FGA. However, all p-values are smaller than 0.05 when their FG%s were tested.

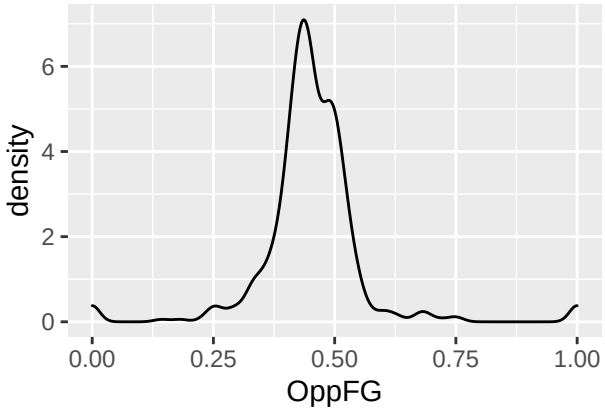
The examination of these three players showed that there is a difference between them and the others, because by raising their shooting accuracy they can more contribute to the winning rates of their teams.

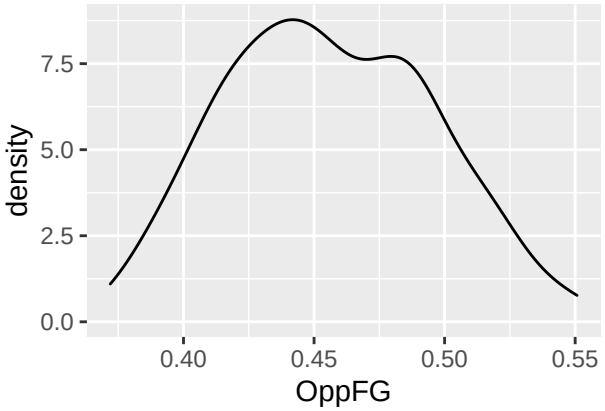
## Analysis on Best Defender

### Best defender

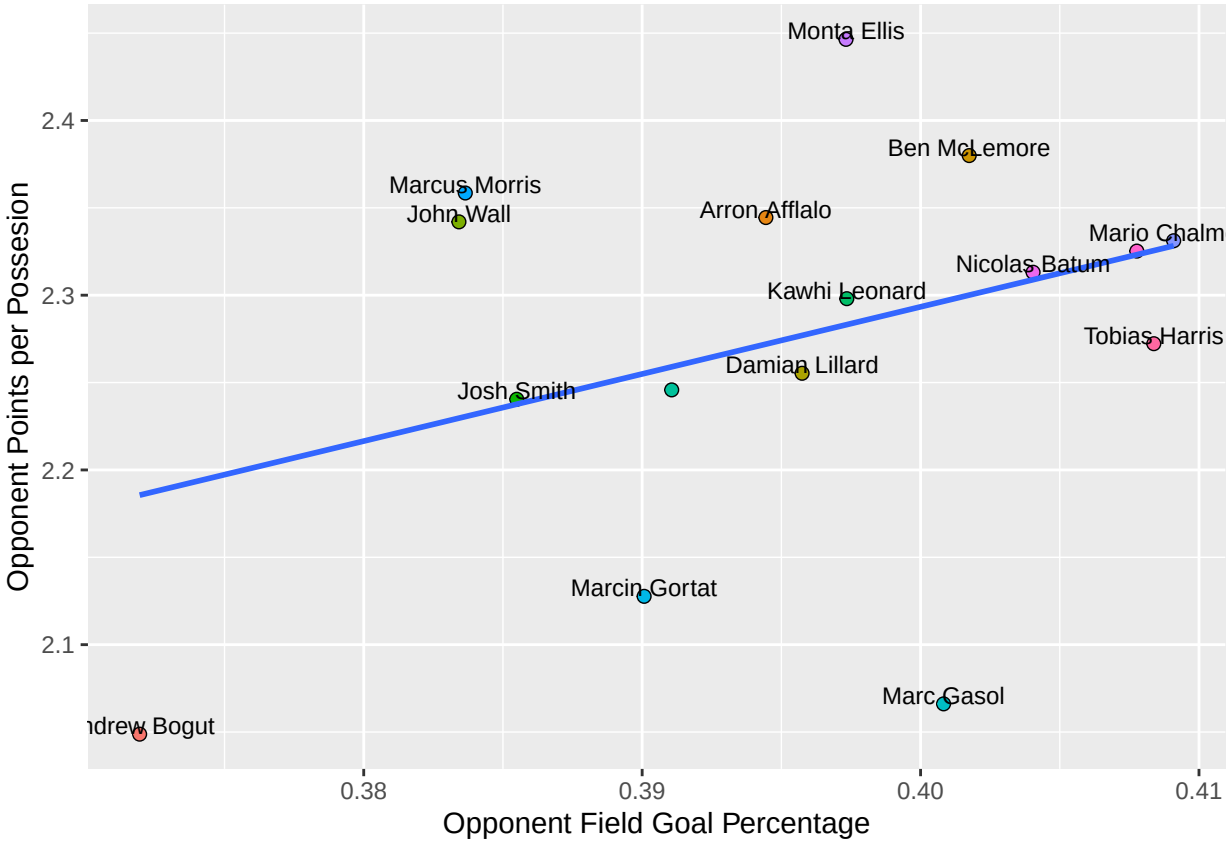
Establish the player defensive ability index data frame DefDF and make statistics:

```
grid.arrange(gg1, gg2,nrow = 2, ncol = 2)
```




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




This gives us the top 15 or so for opponent FG% for defenders who contested 150 or more shots. The list contains mostly players known for their defensive prowess, like Andrew Bogut or Rudy Gobert, but the list also contains players considered average to below-average defensive players in Monta Ellis and Steph Curry. Plotting the two shows a clear correlation between the two statistics, but a great deal of variation still.

## Conclusion

In this report, we have used various statistical tools to explore the NBA data set. We performed some correlation analyses(Pearson correlation) and some tests on means and variances using the t-test and ANOVA, together with various tests for checking conditions. We also examined the top players to compare the result in every case. In addition, we computed and discussed some linear regression and logistic regression models, and used various graphical representation tools to illustrate the data and our findings. For the parametric tests, sample sizes were generally large enough to ensure the validity of the normal approximation framework.

Our main observations can be summarized as follows: 1. The elite basketball players actually cope regularly with the tiredness of their shooting accuracy. As quickly as the game reaches an intensity beyond ordinary tiredness effects, the precision of the shot is reduced in an average. 2. Top players do not get affected by last seconds stress as much as average players. They keep up a quite high accuracy even for the last second of the shot clock (35%). 3. Three individual players: James Harden, Klay Thompson and Monta Ellis are most likely to win in the single match by chart and test when they score more in single game. 4. There was no correlation between the top scorers'

field target attempts and the match outcomes. One probable explanation is that the opponent may focus more easily on the defensive side of a particular player, and the general effectiveness of the team begins to decline. Therefore, overuse of a top player's scoring power is probably not a smart method if the team wishes to enhance the chance of a game. 5.Defender distance in specific ranges has been impacted by the accuracy rate. These observations are in our opinion interesting and appear to be statistically significant. Our conclusions should nevertheless be checked and refined by further analyses, potentially using other sources of data.

## Reference


Investigating NBA Shot Data | Data Science Blog. https://nycdatascience.com/blog/student-works/investigating-nba-shot-data/

https://www.kaggle.com/slangenborg/analyzing-the-best-defenders-in-the-nba/report

Kaggle.com. 2021. Analyzing the Best Defenders in the NBA. [online] Available at: https://www.kaggle.com/slangenborg/analyzing-the-best-defenders-in-the-nba/report [Accessed 12 August 2021].

Ahart (1973). In Kendall, P. and Hollon, S. (1979). Cognitive-behavioral interventions. New York: Academic Press. Economist: As sweet as ever (2015) Available at: http://www.economist.com/blogs/gametheory/2015/06/home-advantage-basketball (http://www.economist.com/blogs/gametheory/2015/06/home-advantage-basketball) (Accessed: 16 October 2016).

Erculj, F. and Supej, M. (2009). Impact of Fatigue on the Position of the Release Arm and Shoulder Girdle over a Longer Shooting Distance for an Elite Basketball Player. Journal of Strength and Conditioning Research, 23(3), pp.1029-1036.

Gilovich, T., Vallone, R. and Tversky, A. (1985) â The hot hand in basketball: On the misperception of random sequencesâ , Cognitive Psychology, 17(3), pp. 295â “314. doi: 10.1016/0010-0285(85)90010-6.

Mascret, N., V., Buekers, M., Casanova, R., Marqueste, T., Montagne, G., Rao, G., Roux, Y. and Cury, F. (2016). The Influence of the â Trier Social Stress Testâ on Free Throw Performance in Basketball: An Interdisciplinary Study. PLOS ONE, 11(6), p.e0157215.