Date prepared: 16/10/2016

Introduction

The purpose of this report is to analyse the chosen dataset and to produce meaningful results using statistical and mathematical analysis. These analyses have been conducted using RStudio software. This report will focused on analysis of a number of areas, demonstrate how the analysis has been conducted, discusses theory used, demonstrate and evaluate produced results. In addition, where relevant, this report will highlight potential areas of future research.

The chosen dataset for this report is a summary of shots made during NBA season 2014-2015. The following areas of research have been selected:

1. Team related
1.1 Home Advantage analysis
1.2 Effect of rest on teams' performance

2. Player related
2.1 Effects on Accuracy
2.1.1 Fatigue Effect (Quarter Accuracy Rate) & Shot Clock Pressure
2.1.2 Defender Proximity Analysis
2.1.3 Shooting Distance vs Shot Accuracy
2.2 Hot Hand Theory
2.3 Game result impact by field goal attempts

Dataset Description:

Data on shots taken during the NBA 2014-2015 season for regular matches throughout the year. The data comprise of 128,609 rows and 22 columns with 16 MB data size with several value in some column missing. The column titles are generally self-explanatory. It has the following data:

Data contains full set of shot attempts by each team/player during the NBA 2014-2015 season for regular matches throughout the year. The data comprise of 128,609 rows and 22 columns with. The column titles (variables) are as follows:
1. Match ID
2. Match date and names of contestants
3. Location: home and away
4. Match result
5. Final score difference
6. shot number
7. FGM: Shot result: success or miss
8. Type of shot attempt (2 points or 3 points)
9. Quarter number (1 to 4, or higher for overtime)
10. Game clock for each quarter when the shot has been made
11. Shot clock (time left for a shot)
12. Name and ID of a player who took the shot
13. Shot distance from basket
14. Name and ID of the nearest defender
15. Distance from the nearest defender

The data has been acquired from Kaggle (https://www.kaggle.com/dansbecker/nba-shot-logs) website on 30 September 2016.

Dataset Strengths

The dataset comprises of comprehensive observational data to start to work with. In general, the dataset has only a small percentage of missing value (0.19%) so there was not need to remove or assign many data points to default values. Since the column name is not ambiguous it was really helpful to understand the dataset right from the beginning.

The dataset comprises of comprehensive observational data which enables efficient analysis without extensive 'clean up' of the data. The dataset only have small percentage of missing value (0.19%), therefore there is no need to remove or assign numerous data points to default values. The fact that initial dataset meant that only a small number of additional columns had to be added (see data clean-up part of the code for details.

Dataset Limitations

There are several limitations of the dataset:
1. The dataset consists only 1 year NBA season stats (2014-2015)
2. The dataset comprise only the data for regular season matches
3. The playoff and other matches during the year are not included 4. The dataset does not include information on free throw shots 5. The dataset does not include final scores of the games

Analysis

1.1 Home Advantage analysis

Introduction

This area of analysis focuses on investigation of whether playing at ‘home’ gives the team advantage comparing to when playing ‘away’. Essentially, this analysis focuses on effect of the game location (home/away) on winning ratio.

The NULL Hypothesis: Analysis of means for home and away wins does not show any significant difference

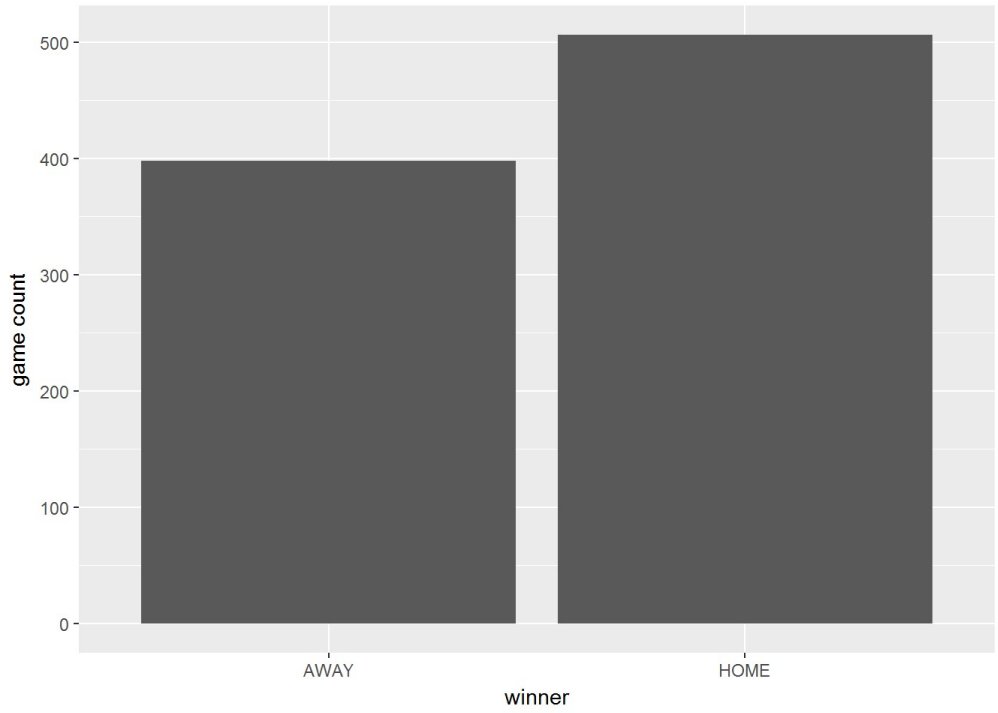
Methods

Steps to test the hypothesis are as follows:
1. Calculate a total number of games, which has been won by a home or away team 2. Compare the total win numbers to identify whether home advantage is evident using means of the variables 3. Extend the analysis to review of the home/away wins on a team level 4. Plot the team level data 5. Perform a t-test (p .05) to examine whether there is a difference in the means between teams winning at home and teams winning away 6. Linear regression with “Final margin”" as dependent variable and “Location” as independent variable to examine whether playing at home predicts a higher final margin

Table 1 below shows descriptive statistics for home/away wins:

Table 1: Proportion of total games split by Location win factor

X	Home Win	Away Win
Game	506	398
Perc	55.97%	44.03%

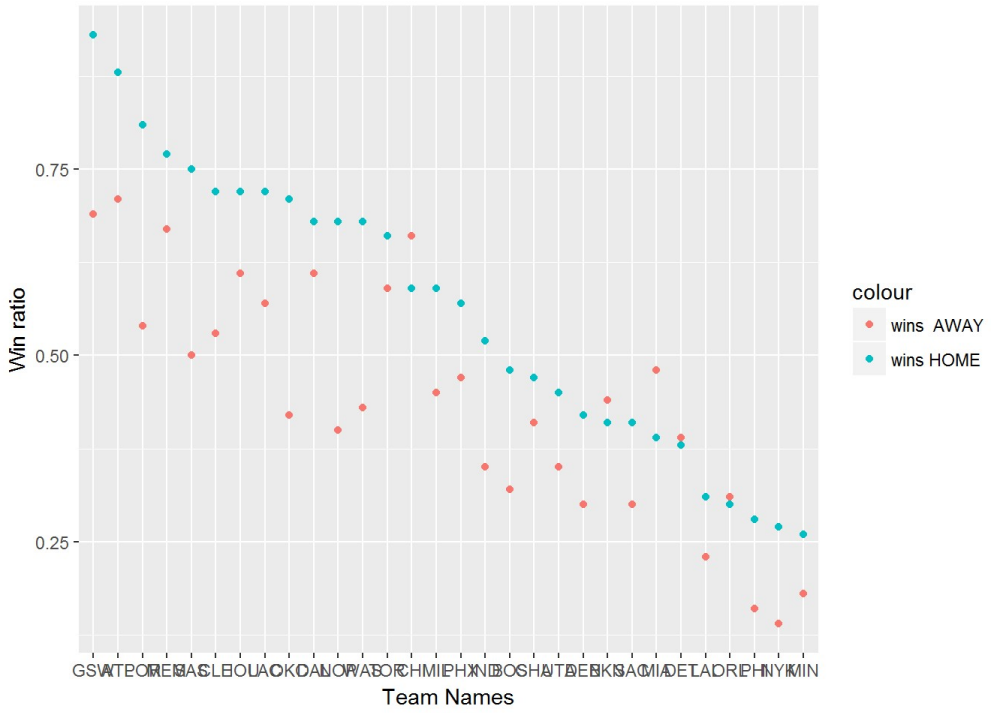


Results

Tables 2 & 3: Winning ratio of teams playing at home vs games playing away

Team	GSW	ATL	POR	MEM	SAS	CLE	HOU	LAC	OKC	DAL	NOP	WAS	TOR	CHI	MIL
Win % HOME	93	88	81	77	75	72	72	72	71	68	68	68	66	59	59
Win % AWAY	69	71	54	67	50	53	61	57	42	61	40	43	59	66	45
Win % Diff	24	17	27	10	25	19	11	15	29	7	28	25	7	-7	14

Team	PHX	IND	BOS	CHA	UTA	DEN	BKN	SAC	MIA	DET	LAL	ORL	PHI	NYK	MI
Win % HOME	57	52	48	47	45	42	41	41	39	38	31	30	28	27	26
Win % AWAY	47	35	32	41	35	30	44	30	48	39	23	31	16	14	18
Win % Diff	10	17	16	6	10	12	-3	11	-9	-1	8	-1	12	13	8



The plot shows that out of 30 teams: 25 teams have won more games at home than away 3 teams have won more games away than at home 2 teams home and away win ratios are almost equal (difference of less than or equal to 1%)

The T-test used to verify the differences in the means seen in the plot reached the level of significance, t(58) = -2.39, p = .02.

Regression analysis

Table 4: Regression analysis for the effect of location and rest days on final score margin

X	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	-2.12919	0.51589	-4.127	3.84e-05 ***
LOCATIONH	4.28556	0.63881	6.709	2.63e-11 ***
rest	-0.01129	0.25992	-0.043	0.965

Adjusted R-squared = 0.02392

Discussion

The analysis shows that there is a clear evidence that a team playing at home is more likely to win with an average win probability of 55.97%. The same conclusion is evident from the analysis conducted on a team level where 25 teams are winning more games at home than away. The results show that the NULL hypothesis should be rejected as a team playing at home is more likely to win than an away team. Linear regression analysis shows that playing at home gives teams on average additional 4.29 margin points.

The result of this research also confirms the recent trend of decline in home advantage over the years: probability of a home team winning has reduced from around 65% in 1975-1992 to average of 60.3% in 1993-2011 and to around 58.5% for 2011-14. The analysis demonstrated that the home win ratio went down even further to 56.0% in 2014-2015 season. Based on the literature review this trend could be potentially explained by the following factors (Economist, 2015):

1. Improvement of travel conditions. This means that when playing away, teams are now more likely to stay at better hotels and fly chartered planes more often than in the past. These changes result in teams being less tired as a consequence of travel and being better physically prepared for the games.
2. Change in style of play in NBA. The game now is more open with less physical contact, which reduces number of fouls and potential impact of ‘notorious home-team bias’ by referees (Economist, 2015).

1.2 Effect of rest on teams' performance

Introduction

This area of analysis focuses on investigation of whether a number of rest days between games affect the performance and more importantly winning ratio of a team.

The NULL Hypothesis is: Teams' mean values of wins are not significantly different based on the number of rest days between games

Methods

The steps to calculate this measurement are as follows:
1. Calculating number of rest days between games for all teams 3. Calculating ratios for effects of rest days on win/loss form home/away 4. Perform a Chi-square test to examine whether there is a difference in the distribution for games won and games lost for the different restdays 5. Calculating ratios for effects of rest days on win/loss form home/away

The table below shows descriptive statistics for Rest Days:

Table 5: Distribution of number of games per rest days

Rest Days	Games	Distribution
0	427	23.62
1	990	54.76
2	254	14.05
3	59	3.26
4	13	0.72
5	3	0.17
7	6	0.33
8	20	1.11
9	3	0.17
10	3	0.17
1st game of season	30	1.66

The distribution of rest days shows that 92% of games are covered by 0,1 or 2 days of rest and therefore further analysis will only focus on these 3 sections.

Results

Table 6: Win rate split by number of rest days between games

Rest Days	Number of Wins	Win Percentage
0	201	47.07
1	496	50.10
2	137	53.94

Table 7: Win rate split by number of rest days between games and location

Rest Days	Home Wins Percentage	Away Wins Percentage
0	53.54	44.33
1	54.68	44.24
2	59.57	46.90

The Pearson's Chi-Square test failed to reach the level of significance, X2 (36) = 42, p= .23.

Discussion

The analysis demonstrates that overall the winning ratio stays the same with increase in rest days. However, when looking at individual restdays the findings are different: back-to-back games have 47.07% winning ratio, 1 day rest increases winning ratio to 50.1% and with 2 days rest the winning ration increases even further to 53.94%.

By extending the analysis to the team level, the results demonstrate similar tendency for home games where winning ratio increased from 53.54% to 54.68% and 59.57% respectively with 0, 1 and 2 days of rest. Somewhat interestingly the analysis of effect of rest days on winning ratio of away games demonstrates that there is hardly any difference between 0 and 1 days of rest (44.33% and 44.24%). However with 2 days of rest teams' average winning ratio increases to 46.90% (2.66% increase). This perhaps could be explained by the fact that teams were able to minimise effect of travelling when having a bigger window between games.

Based on the above, the analysis shows that the null hypothesis should be rejected as there is a clear evidence that the higher rest days correspond to higher winning ratio.

Further research

To better understand effects of rest and home/away advantage future research should focus on distances the teams have to travel between the games. For example if a team is travelling from one coast to another- it is likely to affect them more comparing to a team from New York 'travelling' to play an away game in New Jersey.

2.1 Effects on accuracy

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

Introduction

The potential effects of fatigue and stress on the accuracy of shots are investigated in this part of the analysis.

Fatigue in basketball is a broadly analysed topic in physiological research. Latest literature suggests that elite basketball players cope well with the effects of fatigue so it does not affect their shooting kinematics. This results in no difference for their shot accuracy even if the players are fatigued (Erculj and Supej, 2009). This research paper analyses whether these previous findings can be confirmed for the NBA season 2014-2015.

Regarding the effect of psychological stress on shot accuracy, previous research is not as decisive as for the effects of fatigue. While Ahart (1973, in P.C. Kendall & S.D Hollon (1979) *Cognitive-Behavioral Interventions: Theory, Research, and Procedures*) concluded that stressful events like small or big margins for the game score have an effect on players free-throw accuracy, latest research introducing more controlled conditions states that there is no effect of stress on players shot accuracy (Mascret et al., 2016). The indecisiveness of previous literature led to a deeper analysis in this research paper.

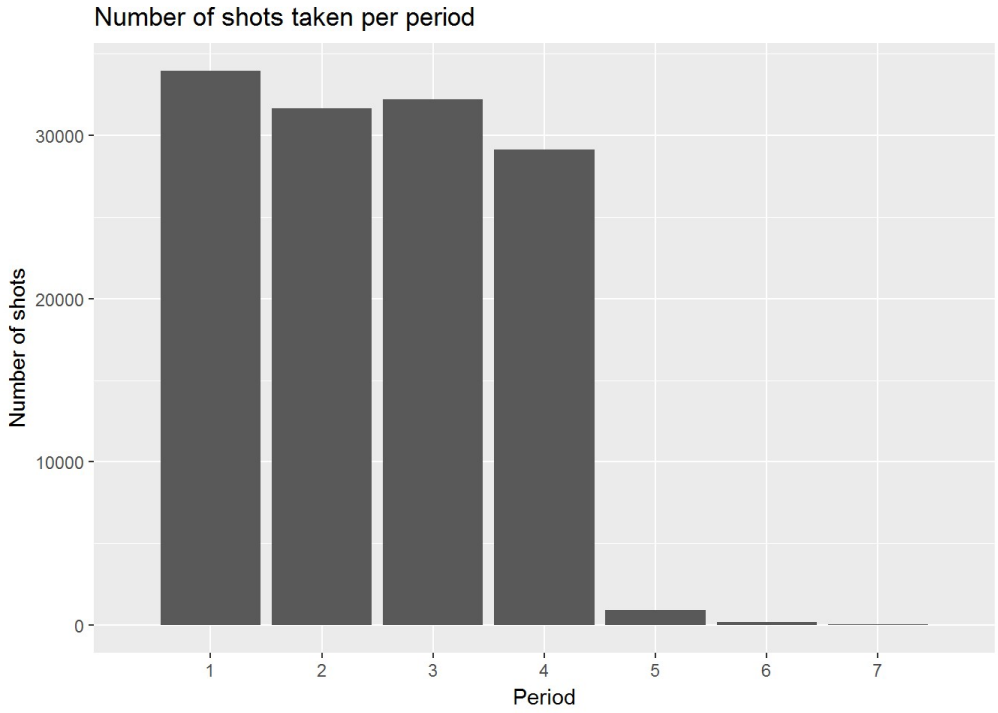
Hypotheses for this area of analysis are:

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.

Methods

In this analysis fatigue was measured as the period of the game, as it can be assumed that the later in the game the higher the fatigue of the players. The descriptive statistics of the shots per period can be seen below.



The steps to test the first hypothesis were as follows: 1. The overall accuracy for each period was calculated, based on the number of shots made (FGM = 1) and the number of shots taken for the specific period (FGM = 1 & FGM = 0) 2. This results in a percentage value of shots made for each of the periods 3. A variable called “Overtime” was introduced to distinguish between periods in the regular game time and overtime 4. The average accuracy for regular time and overtime was calculated 5. A T-test was used to examine whether or not these values were significant different at a level of significance of 5%.

Psychological stress was measured through the remaining time on the shot clock for each shot taken. It was assumed that a player would experience more stress the lower the shot clock goes.

The steps to test the second hypothesis were as follows: 1. As the shot clock variable contained 5567 missing values, those were excluded from the analysis 2. To simplify the shot clock variable, decimals were removed from the data so every shot taken was assigned to 0 to 24 seconds left on the shot clock 3. The overall accuracy for each second was calculated using the FGM variable

Results

The accuracies for each period are detailed in the table below.

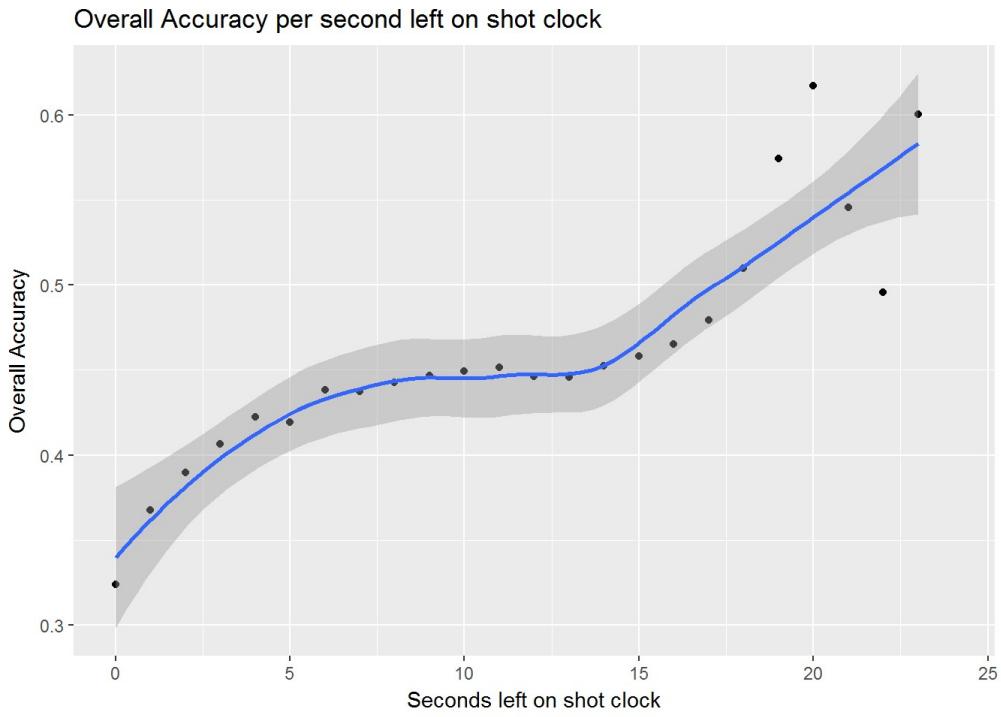
Table 8: Accuracy per period

Period	Accuracy
1	46%
2	45%
3	46%
4	44%
5	39%
6	43%
7	37%
regular	45%
overtime	40%

The T-tests for period 1 and period 2 was significant, t(65301) = 2.42, p = .015, as well as the T-test for period 3 and 4, t(60757) = 4.24, p < .01. Further the T-test for period 2 and 3 was not significant, t(63843) = -1.53, p = .126, as well as the T-test for period 1 and 3, t(65989) = 0.87, p = .38. The T-test for regular time and overtime was highly significant, t(383) = 2.77, p < .01.

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

```
## `geom_smooth()` using method = 'loess'
```



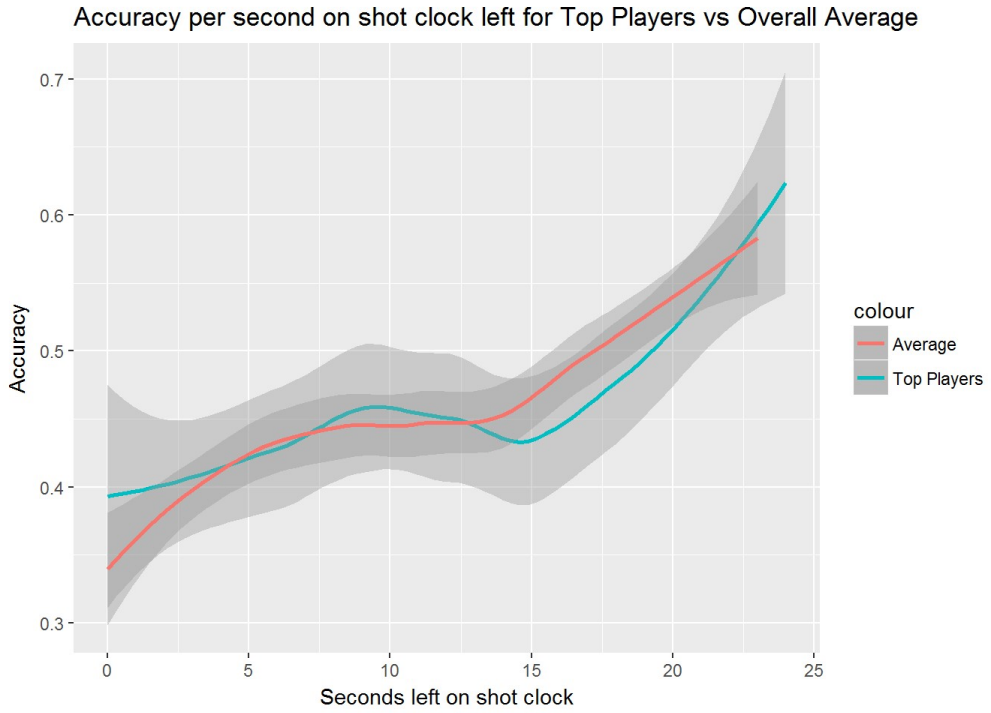
Discussion

The results of the T-tests show a significant difference for the overall accuracy between period 1 and period 2, between period 3 and period 4 and between regular game time and overtime. This means that player have a higher accuracy in periods 1 and 3 than in the following one. As the table 8 shows, the changes in accuracy are minor (from 46% to 44% over the regular game time) and often located on digit-level, so although the changes are significant they are negligible. After all, the Null hypothesis must be rejected as the statistical methods used show differences for accuracy in later periods of the game, but as they are negligible in the regular game time but it can be concluded that accuracy does not suffer for later game periods as long as the game ends in regular time. For games that reach overtime a highly significant drop in the accuracy of over 5% (45% to below 40%) is observable, stating that players are less likely to score in overtime periods that in the regular game time. This means that elite basketball players cope really well with any occurrence of fatigue regarding their shot accuracy on a regular basis. As soon as the game reaches an intensity which is beyond the regular level effects of fatigue can be found leading to lower shot accuracy.

The effects of psychological stress are highly visible as shown in the graph above. While the accuracy in the first 10 seconds of the shot clock is exceeding the overall accuracy in the regular game time (60-45% vs 45% in regular game time), it declines quite fast in the last 5 seconds of the shot clock (from 42% to 32%), leading to the rejection of the Null hypothesis, that there are no differences in the accuracy for more stressful events. These results imply that as the shot clock runs out, so as the psychological pressure on the shooter increases, their shot performance suffers.

An ad hoc data analysis was undertaken to investigate whether this pattern of stress-inducted accuracy declination can be found as well for “superstar” players. The players were Lebron James, James Harden and Russel Westbrook, chosen because they were the top 3 shooters in the season. Their combined average accuracy per second can be seen in the graphic below.

```
## `geom_smooth()` using method = 'loess'
## `geom_smooth()` using method = 'loess'
```



The graph shows many similarities for top players to the average graph. The most visual differences are the slightly lower accuracy around the 15 seconds mark for the top players and the much lower decrease below 5 seconds left. Explanations for these differences could be that around the 15 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 accuracies. Further it seems like Top players don't 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 (40%). However, as there are many similarities for the accuracy of Top players and average players there must be other things than the accuracy that separates those two groups.

Future research

Future research could focus on this difference between Top players and average players. While this analysis revealed a difference for shot clock pressure it would be interesting to look at more events that induce stress like the margin of the game or the placement in the standings of the own and the enemy team. Further an investigation of the reasons that lead to the higher accuracy would be interesting. Do top players keep up their usual shot kinematics when under pressure while average players don't? Is it a different mindset that helps them to cope better with stress? Or do they simply have a higher stress tolerance as the result of continuous high expectations top players have to endure?

Additionally, further reasons that lead to distinction between top players and average players should be investigated.

Limitations

As there are much more shots taken in periods 1 to 4 than in 5 to 7 the differences in the accuracy have to be treat with caution. The shot clock variable had missing values for 5567 shots, equal to 4.34% of all shots on the dataset, limiting the meaningfulness of the findings in the accuracy per second part.

2.1.2 Closest Defender Distance vs Shot Accuracy

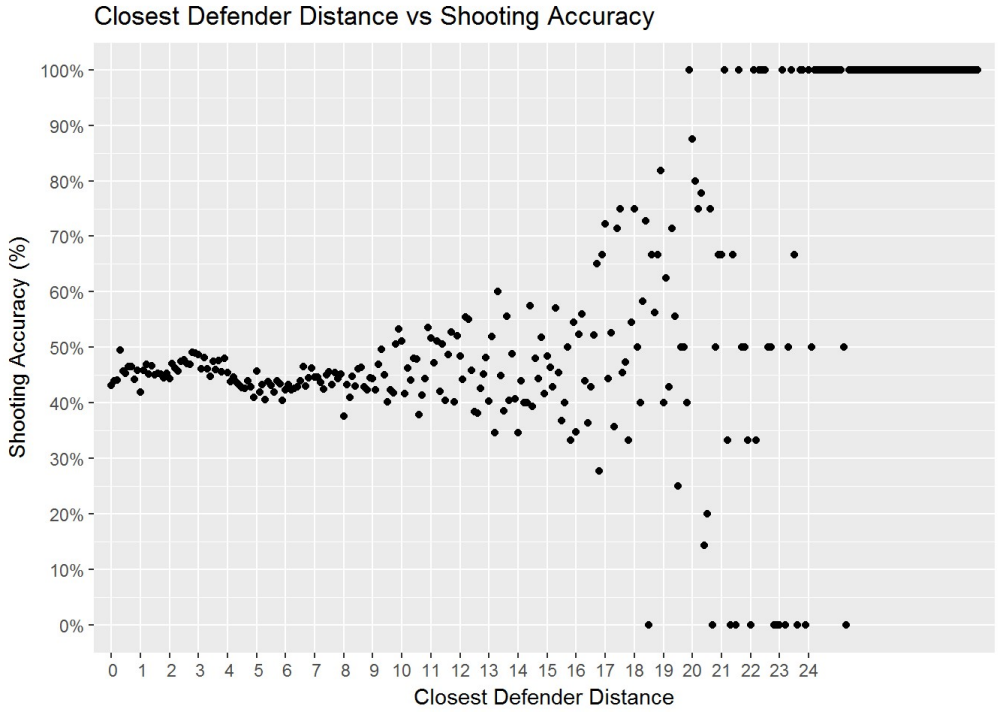
Introduction

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.

Methods

The steps to calculate this measurement are as follows:
1. Assign shot result to determine whether the shot is scored (1) or missed (0)
2. Select and analyse subset of all closest defender distance values
3. Count the occurrence for every distance
4. Calculate a number of successful shots for every distance point
5. Calculate accuracy ratio for every defender distance value 6. Separate data in two ranges - 0 to 8 feet distance and above 8 feet distance



Results

Plot Result

Table 9: Relationship between the defender distance and accuracy

Defender Distance (range in feet)	accuracy (range in %)
0 - 2	42 - 50
2 - 4	44 - 49
4 - 6	41 - 46
6 - 8	37 - 42
8 - 10	37 - 54
10 - 16	33 - 60
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 Pearsonâs correlation coefficient for the relationship between the variables “Defender Distance” and “Accuracy” was r(297) = .59, p < .01, with a level of significance of p = .05.

Correlation Test Result

When analysing 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 medium negative correlation between the defender distance and shooting accuracy in this range. The Pearsonâs correlation coefficient for the relationship between the variables Defender Distance and Accuracy was r(297) = -0.40, p < .01, with the level of significance of p = .05.

Discussion

The analysis demonstrated that the accuracy rate was affected by the defender distance for certain ranges. For the defender distance starting from 0 to 8 feet, there is a negative correlation between distance and accuracy, implying that players are more likely to miss when the defender comes closer. For the distance greater than 16 feet, the data varies between 0% (no shots scored from that distance) to 100% (certain successful shot from that distance), which means that this part of the analysis should not be taken into account as the results were insignificant. 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.

Future Research

As the dataset did not hold many data points for more than 20 feet defender distance, the accuracy analysis for these distances are limited. The future analysis could be enhanced further by including data for additional seasons.

2.1.3 Shooting Distance vs Shot Accuracy

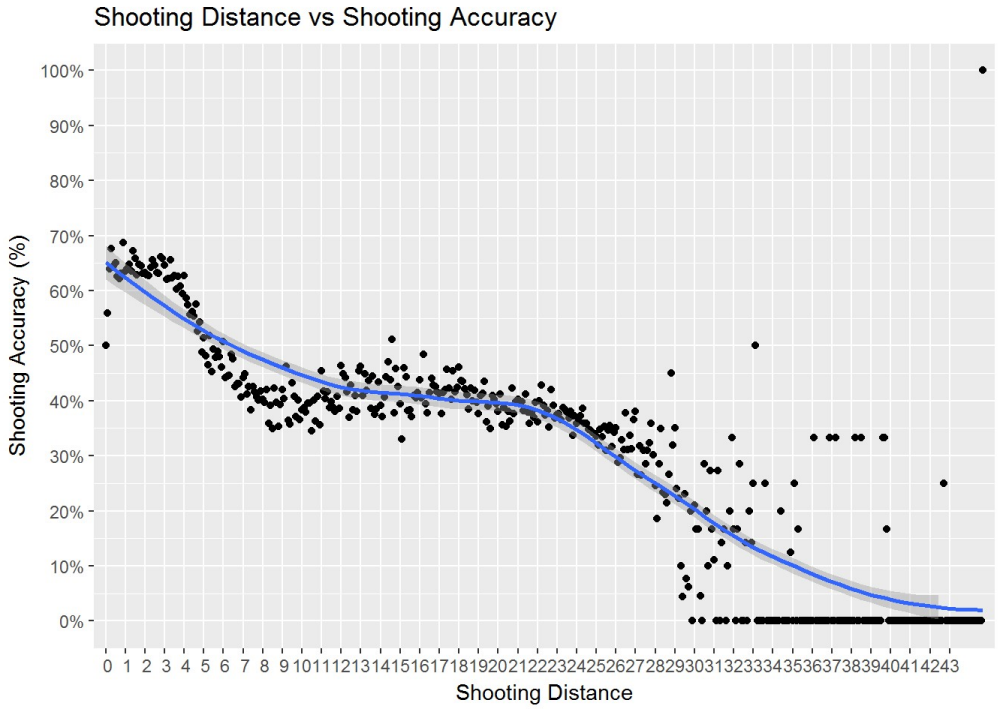
Introduction

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.

Methods

The steps to calculate this measurement are as follows:
1. Assign shot result to determine whether the shot is scored (1) or missed (0)
2. Select subsets of all shooting distances from the basket
3. Count the occurrence for every distance to basket value
4. Allocate shot success rate to every distance from basket value 5. Examine the relationship between distance to the basket and shot outcome with a correlation test 6. Separate data into ranges



###Results

Plot Result

Table 10: The relationship between the shooter distance ranges and accuracy

Distance (in feet)	Accuracy (range in %)
0 - 2	62 - 68
2 - 4	61 - 67
4 - 6	44 - 58
6 - 8	38 - 51
8 - 10	35 - 42
10 - 12	34 - 42
12 - 14	37 - 47
14 - 16	33 - 51
16 - 18	37 - 48
18 - 24	33 - 43
24 - 28	18 - 38
28+	0 - 50

There is a strong negative correlation between the shooting distance and shooting accuracy. The Pearsonâs correlation coefficient for the relationship between the variables Shooting Distance and Shot Accuracy was r(446) = -0.87, 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 Pearsonâs correlation coefficient for the relationship between the variables Shooting Distance and Shot Accuracy was r(99) = .11, p =.24, failing to the level of significance (p = .05).

Discussion

Based on the diagram and correlation above, it could be concluded that although the shooting distance affects the shooting accuracy, the relationship between these variables is not perfectly linear. This was shown in the diagram where the accuracy for shots taken from 10 to 20 feet is relatively similar, regardless the shooter's performance. Regarding the low success rate for higher distances, it could be arguably inferred that the shooter tried to attempt a shot in closing game time (buzzer beater) or when the shot clock or game clock nearly reach zero - Therefore, the analysis would reject the Null Hypothesis.

Future Research

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.

2.2 Hot Hand Theory

Introduction

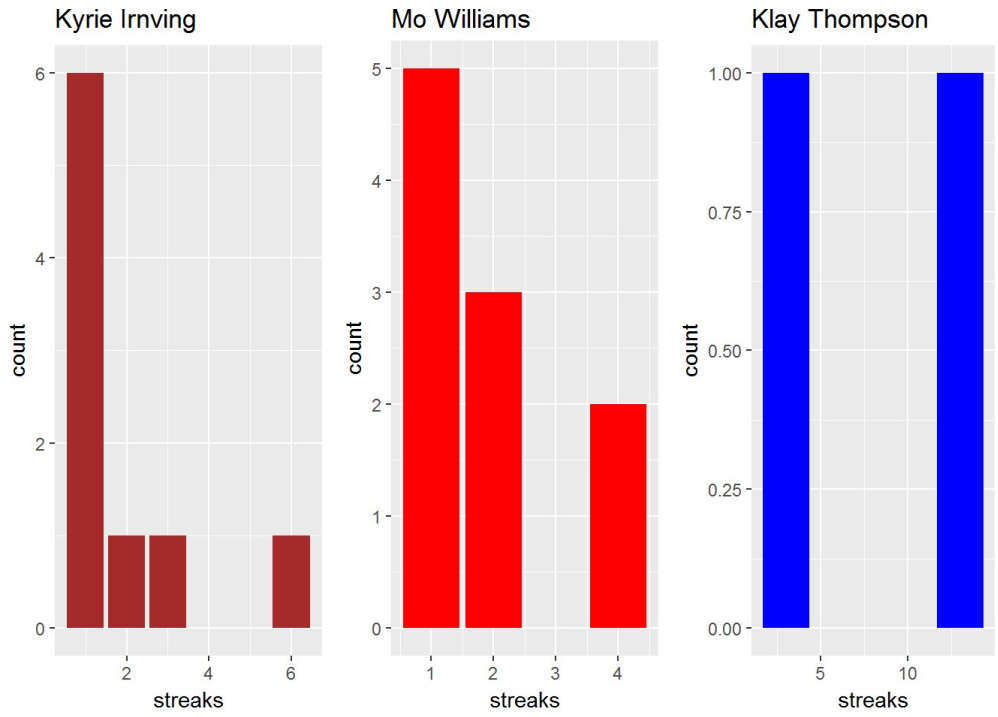
The analysis aims to investigate whether the effect of the 'hot hand theory' is valid based on the given dataset. The 'hot hand theory' states that if the shooter makes his first shot, the probability of making his next shot is higher. (Gilovich et al., 1985)

Methods

The steps to approximate this question are as follows: 1. Calculate streaks achieved in all games and from all players 2. Categorize streaks per player and game, focusing on the 3 players with the top 3 performances in terms of points 3. Extract zeros and keep only streaks above 1 4. Visualise streaks' length distribution - barplots 5. Visualise streaks' length via boxplot 6. Calculate the shooting percentage of the 3 players 7. Create a sample independent shooter 8. Compare distributions of sample and real player

Results

Barplots



Barplot result

The distribution of the three players is unimodal and left skewed. There are shown some extremely long shooting streaks made (i.e. Klay Thompson with 13 shots in a row).

Boxplots

Outcome for Kyrie Irnving:

1. The typical length of a streak is 1
2. The Interquartile Range is 1
3. Streak length of 6 is unusually high compared to the rest of the distribution

Outcome for Mo Williams

1. The typical length of a streak is 1.5
2. The Interquartile Range is 1
3. Streak length of 4 is unusually high compared to the rest of the distribution

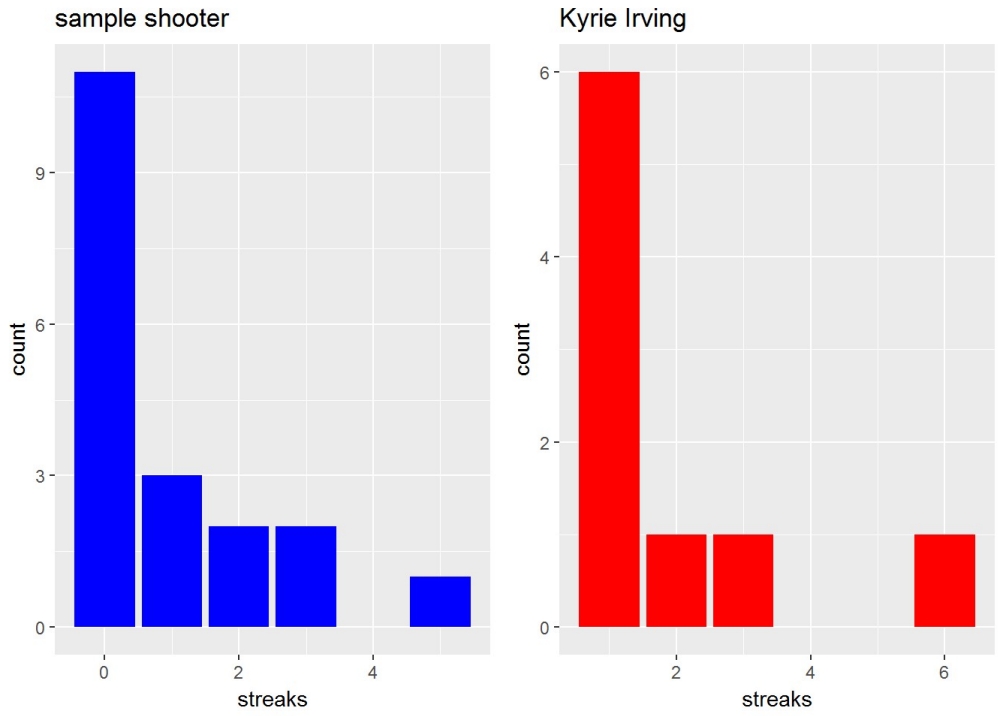
Outcome for Klay Thompson

1. The typical length of a streak is 8
2. The Interquartile Range is 5
3. There is a streak of length 13

Plots Result

The distribution of all players is left skewed and all of them (especially Klay Thompson) have some long shooting streaks. In order to prove/disprove the hot hand theory, the examination of the independence of the shots is required. If each shot that a player takes is independent of the next shot, then the player has the same probability of hitting each shot regardless of the previous shot. If each shot that a player takes is dependent of the next shot, then the probability of making the next shot is higher - hot hand. Sampling an independent shooter having the same shooting percentage with the real one and comparing his distribution with real player's distribution gives further insight in the specific area.

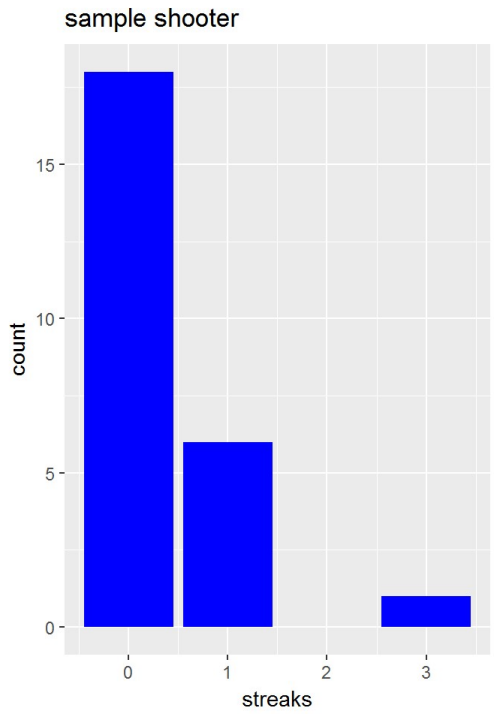
Kyrie Irving

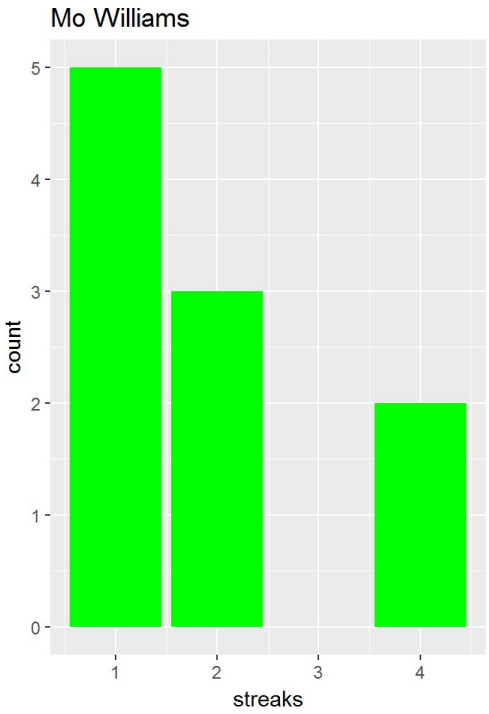


Plot Result

After producing a sample of shots for a player, the result shows distributions of sample and real shooter are quite similar. As a consequence, it can be concluded that Kyrie Irving does not have a hot hand.

Mo Williams

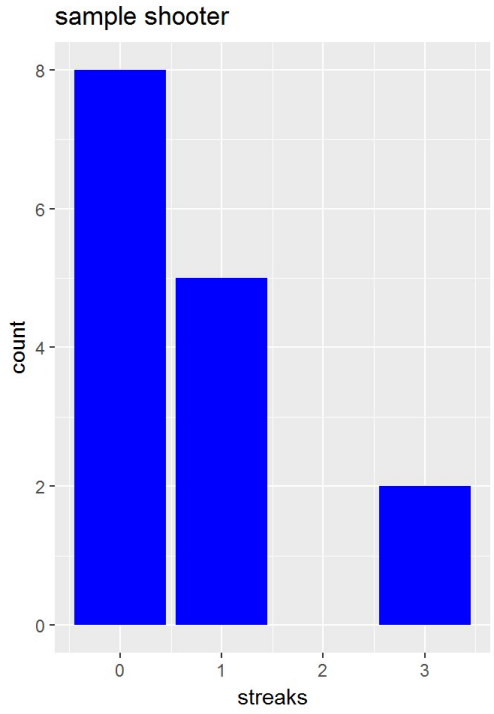


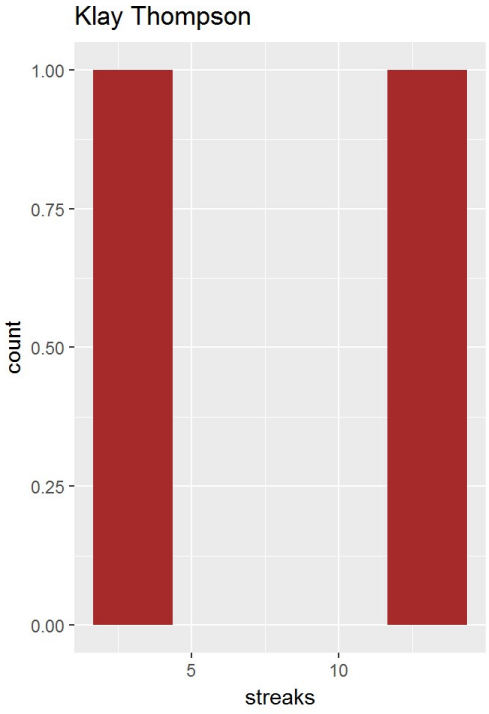


Plot Result

After producing a sample of shots for a player, the result shows distributions of sample and real shooter are quite similar. As a consequence, it can be concluded that Mo Williams does not have a hot hand.

Klay Thompson





Plot Result

Klay Thompson is an exception making 13 shots in a row. There is no such distribution in the sample, so it could be assumed that he is likely to have a hot hand. However, further analysis of Klay Thompson's (and other basketball players) shooting streaks in other games is required to see if such a rare occurrence is repeated(this is out of scope for this analysis).

Discussion

The analysis demonstrates that the streaks generated by the players (K. Irving & M. Williams) are similar to randomly generated streaks, which use these players shooting probabilities. As a result there is no evidence to suggest that there is a hot hand phenomenon present here.

In the case of K. Thompson, his streak is seem as somewhat anomaly and further research is required to see if any similar results have been achieved by the player.

Limitations

Since the dataset is referred to only one season, the size of the vectors of streaks for each player is very small, a fact that can lead to misleading results. Moreover, streaks do not constitute continuous variables and as a result they cannot be used in a linear regression.

Future research

Since the dataset is referred to only one season and the analysis is not able to eliminate the hot hand theory, it would be an interesting topic to be investigated by using data from more than one season.

2.3 Game result impact by field goal attempt

Introduction

The aim of this analysis is to examine whether the top scorer of a team can help his team to win the match when he has a more outstanding performance in terms of both number of shooting attempts and accuracy. The Null hypothesis is that there is no difference regarding the likelihood of winning if the top scorer of a team has more field goal attempts. The second Null hypothesis is there is no difference regarding the likelihood of winning if the top scorer of a team has a higher shot percentage than on average. As this is such a niche area research a literature review did not reveal sufficient results so this analysis aims to fill this gap.

Methods

The measurements are calculated in the following steps:

1. âMadeâ and âmissedâ in âFGMâ column were replaced by 1 and 0 respectively.
2. A new subset was created so that each row contained the playerâs name, match details and shooting performance for that player.
3. The number of attempts and field goal percentage of all players were calculated according to the shooting data.
4. The players with most outstanding performances were picked out and another subset was generated for these players.
5. The correlation (p = .05) between the results of the matches and the performances regarding FGA and FG% of top scorers was investigated.
6. A logistic regression analysis with the outcome of the match as dependent variable and FGA and percentage of FG made as independent variables was conducted.

Impact of Top Scorers

Results

a. 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 Pearsonâs correlation coefficient for the relationship between the variables W and FGA was r(832) < .01, p = .794, failing to reach the level of significance (p = .05).

b. 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 Pearsonâs correlation coefficient for the relationship between the variables W and FGA was r(832) < .01, p = .177, failing to reach the level of significance (p = .05).

Table 11: Result for the logistic regression analysis for FGA and FG% of 15 players

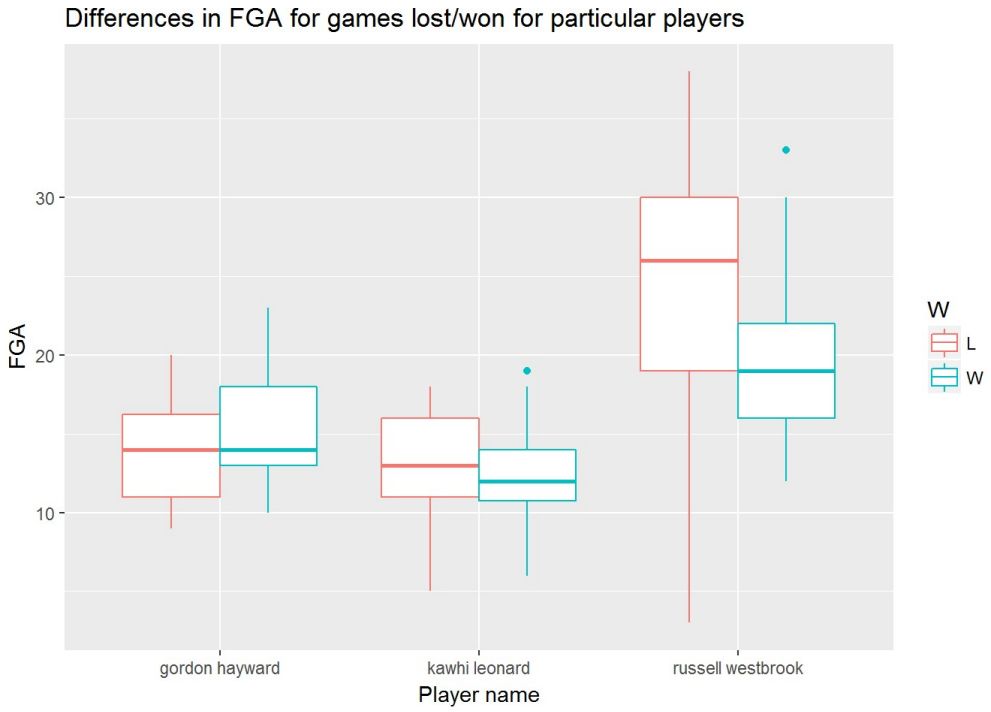
	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	0.042259	0.164816	0.256	0.798

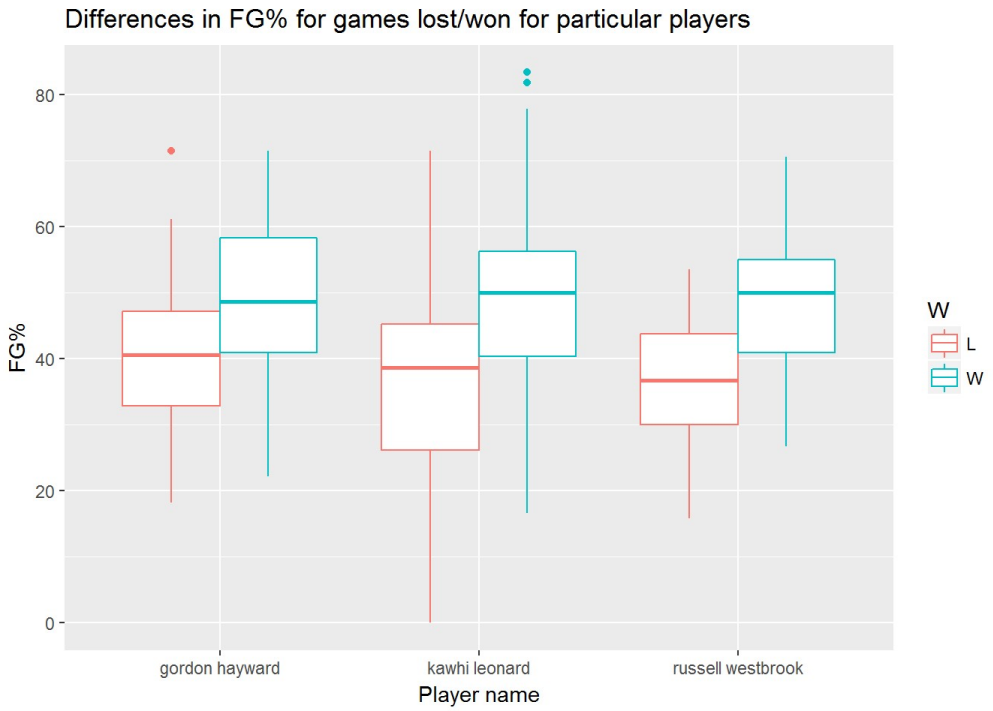
	Estimate	Std. Error	t value	Pr(>t)
FGA	-0.002679	0.006468	-0.414	0.679
FG.percentage	0.012329	0.002736	4.506	1.35e-05 ***

Table 12: Result for the anova for FGA and FG% for 15 players

	Df	Deviance	Resid. Df	Resid. Dev
NULL			777	188.73
FGA	1	0.0014	776	188.73
FG.percentage	1	7.4664	775	181.26

An ad hoc 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: Russell Westbrook, Kawhi Leonard and Gordon Hayward.





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

Table 13: Result for the logistic regression analysis for FGA and FG% of the 3 special players

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	0.042259	0.164816	0.256	0.798
FGA	-0.002679	0.006468	-0.414	0.679
FG.percentage	0.012329	0.002736	4.506	1.35e-05 ***

Table 14: Result for the anova for FGA and FG% for the 3 special players

	Df	Deviance	Resid. Df	Resid. Dev
NULL			147	36.669
FGA	1	0.0498	146	36.619
FG.percentage	1	4.4978	145	32.121

Discussion

From the correlation tests between the match results and playersâ 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 the logistic regression analysis for FGA and FG%, it was found that FGA did not have any impact on the result of the match. FG% is a significant predictor for the result of match. However, the anova showed that the residual deviance of these two factors were both quite large and thus further investigations must be done in order to improve this model.

There was no correlation found between field goal attempts of top scorers and outcome of the matches. A possible explanation is that it is easier for the opponent to focus on a certain player on defensive side and then the overall efficiency of the team will start to decrease. Thus, over-using the scoring ability of a leading player is not likely to be a good approach if the team wants to increase the likelihood of winning a game. Meanwhile, the accuracy of these players has a stronger impact on the outcome of the game. This could be explained by that once a player is being efficient on the offensive side, that player can help the team not only by shooting but also by creating more opportunities for his teammates.

Based on the results above, an additional analysis was done regarding the performance of three particular players: Russell Westbrook, Kawhi Leonard and Gordon Hayward. These three players were chosen because their teams are more likely to take the victory when they score more in single game according to Diagram and t-test.

For these three players, 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. Based on these findings, another logistic regression was done with these three players. It was shown that p < .05 and residual deviance from anova dropped by 4.498 which is about 12.3% when FG% was considered. In comparison, the deviance was above 180 when 15 players were analysed. For these three players, their teams are more likely to win by 1.2% when their FG% increase by 1.

Therefore, the analysis on these three players has shown that there is a difference between them and the others as they can contribute more to the win rates of their teams by their increasing shooting accuracy. There could be various reasons behind this observation, such as their team compositions and the roles of these players. Further research has to be done to in order to find out why the performances of these special players are more decisive on the outcome.

Limitations

The data set only covers 904 of 1230 matches in the regular season. So there might be a chance that the highlight performances of certain players are missed out in the analysis. Meanwhile, it would be ideal if more aspects of data, such as rebounds and assists, could be analysed.

Conclusion

This report has demonstrated results of mathematical and statistical analysis of the dataset containing information of shots from NBA season 2014-2015. The report specifically focused on number of research areas covering teams and players performance. The report described steps taken to produce the results, discussed and evaluated the produced results and suggested potential areas for future research.

References

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) âThe 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., IbÃ¡Ã±ez-GijÃ³n, J., BrÃ©jard, V., Buekers, M., Casanova, R., Marqueste, T., Montagne, G., Rao, G., Roux, Y. and Cury, F. (2016). The Influence of the âTrier Social Stress Testâ on Free Throw Performance in Basketball: An Interdisciplinary Study. PLOS ONE, 11(6), p.e0157215.