An exploratory analysis of the NBA datasets

Chloe Powell (C23041975) and Amina Mannan (C23031410)

Abstract

In this paper, we investigate the 2014 NBA season dataset using statistical techniques including ANOVA, Welch's t-test, Chi-squared test and logistic regression. We examine variable relationships and assess the impact of home advantage.

Contents

1 Introduction

The data used within this report is from `NBA` and `Player_Info`. The `NBA` data set consists of game activity for the 2014 to 2015 NBA season and `Player_Info` data set contains personal information of the players during the same NBA season. The National Basketball Association is the top professional basketball league in the world, based in the United States and Canada, the top performing team from 2014 to 2015 was Golden State Warriors. One can refer to the NBA Wikipedia page for more information: https://en.wikipedia.org/wiki/National_Basketball_Association.

The `NBA` data set consists of 124,364 observations and 21 variables. The `Player_Info` data set consists of 475 observations and 13 variables. The names of the variables are listed below:

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

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

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

```
##  [1] "Age"         "Birth_Place" "Birthdate"   "College"     "Experience"
##  [6] "First_Name"  "Height"      "Pos"         "Surname"     "Team"
## [11] "Weight"      "BMI"         "PLAYER_NAME"
```

2 Variable Relations

In this section, we explore notable relationships between key variables from the `NBA` and `Player_Info` data sets. This analysis leverages statistical techniques to highlight meaningful trends and correlations across the data set's. In Section 2.1, we analyse effects of player height in decision making by assessing dribbles before a shot taken, using ANOVA and Tukey comparison in Section 2.1.1. In Section 2.2 we consider the 20 best and worst players according to shot accuracy, using Histograms and Confidence Intervals we identify trends and in Section 2.2.1 we do further analysis of the findings from Section 2.2. All analysis conducted in Section 2 is influenced by the correlations shown in Figure 1.

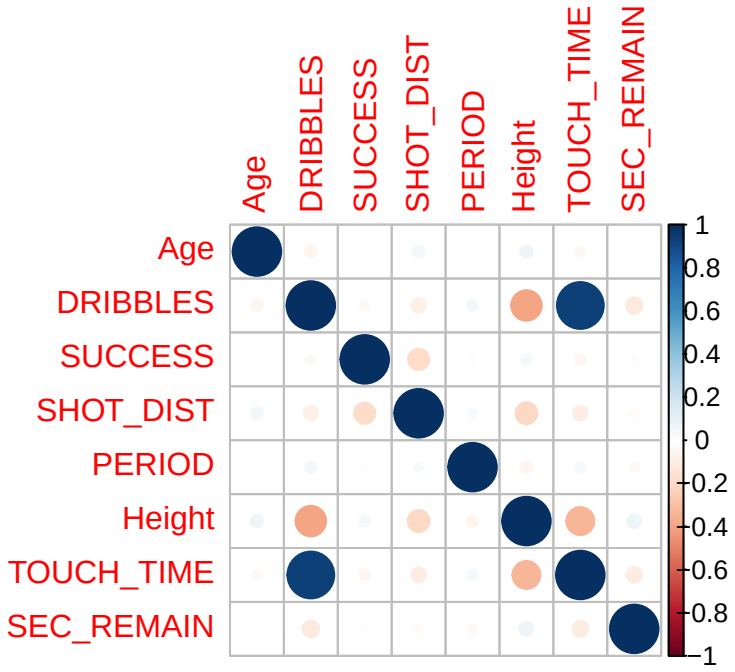

Figure 1: Pearson correlation of variables analyised within this section

2.1 The Impact of Height

In Figure 1 we are able to see the negative correlation between `DRIBBLES` and`Height`, Pearson Correlation Coefficient ≈ -0.3942 . From this observation, we will do further analysis to test whether the hypothesis that the height

of players affects decision-making against the null hypothesis that it has no effect. Therefore, we conducted ANOVA. We can see below in Section 2.1.1 that the F-value is very large, $F = 1425$, indicating that the variation of dribbles between height groups is substantially larger than the variability of dribbles within height groups.

2.1.1 ANOVA

```
combined_data <- merge(NBA, Player_Info, by = "PLAYER_NAME")
combined_data$Height <- factor(combined_data$Height)
res.aov1 <- aov(DRIBBLES ~ Height, data=combined_data)
summary(res.aov1)
```

```
##                 Df  Sum Sq Mean Sq F value Pr(>F)
## Height          16  234416   14651    1425 <2e-16 ***
## Residuals   108109 1111796      10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From these findings we next do a Tukey multiple comparisons test to identify which height groups have statistically significant differences.

```
TukeyHSD(res.aov1)
```

Table 1: Summary of the Tukey comparrison of mean dribbles within height groups.

Height	diff	lwr	upr	p_adg
212.5-195	-1.8550621	-2.5313418	-1.1787823	0.0000000
182.5-177.5	0.5564984	-0.4548465	1.5678433	0.8944546
205-180	-4.6903392	-5.3964368	-3.9842415	0.0000000
190-185	0.1362598	-0.4976657	0.7701854	0.9999983
215-212.5	0.3803867	-0.7923146	1.5530881	0.9995471
210-172.5	-3.9103099	-5.0377280	-2.7828919	0.0000000

In Table 1 we have compiled an overview of the results obtained from the Tukey test. It is made evident in the Table 1 that there is a significant difference in the lower end of height with the upper end of player height as the $p - values < 0.05$ suggest a statistically significant difference as we expected. Therefore there is strong evidence to reject the null hypothesis from Section 2.1.

2.2 The Influence of Dribble Counts on Shot Success Rates

Another interesting negative correlation we can observe in Figure 1, is between `DRIBBLES` and `SUCCESS`, Pearson correlation coefficient ≈ -0.03532. We can preform various tests to see whether this is statistically significant and if there is a trend of dribbles among the NBA's best preforming players. To view the full list of which players are within top 20 and bottom 20 players in accordance to their shot accuracy we can see Table 3 in the Appendix, Section 5.

We create two histograms to visualize the distribution of the data seen in Figure 3. The histograms are presented in Figure 2 along with the 99% confidence intervals. It is evident in the left histogram in Figure 2 that the top 20 players from the season tend to do more dribbles before a shot than those in the bottom 20 players. To test this hypothesis we need to conduct further statistical tests, we can see this in section 2.2.1.

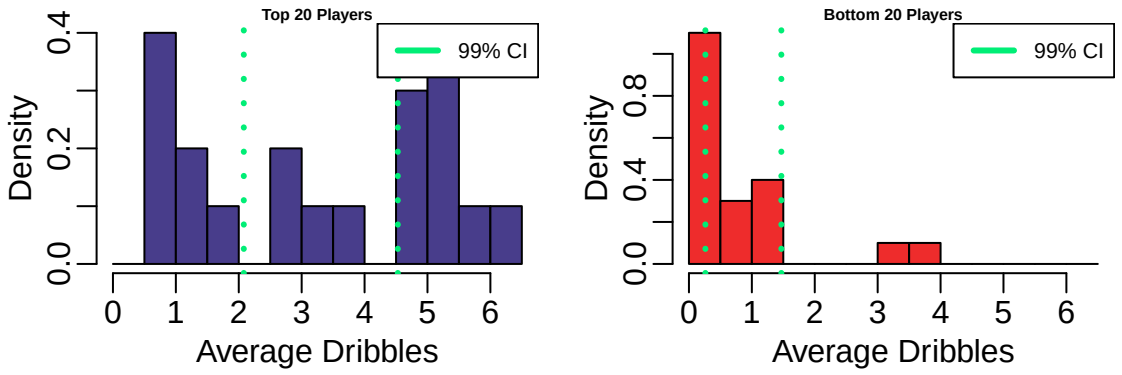

Figure 2: Histogram of the mean dribbles before a shot of the top and bottom 20 players according to shot accuracy for the season and 99% confidence intervals.

2.2.1 Further tests

Here we have conducted a variance test and a t test to check our hypothesis that players with a higher shot accuracy take longer before shots against the null hypothesis that there is no difference between groups.

- Firstly an F-test to compare variances is conducted with significance level $\alpha = 0.05$. This test results in $F - value \approx 4.12$, and $p - value = 0.0034$.
- Secondly a Welch two sample t-test for `top_player_accuracy` and `bottom_player_accuracy` with significance level $\alpha = 0.05$. We see that we're 95% confident that the difference in average dribbles is between $[1.46, 3.42]$ and we have a $p - value \approx 0.000021$.

```
##
##  F test to compare two variances
##
## data:  top_player_accuracy$ave_dribbles and bottom_player_accuracy$ave_dribbles
## F = 4.1157, num df = 19, denom df = 19, p-value = 0.003385
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   1.629033 10.398043
## sample estimates:
## ratio of variances
##           4.115672
```

```
##
##  Welch Two Sample t-test
##
## data:  top_player_accuracy$ave_dribbles and bottom_player_accuracy$ave_dribbles
## t = 5.1123, df = 27.718, p-value = 2.097e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.461387 3.416917
## sample estimates:
## mean of x mean of y
## 3.3041651 0.8650129
```

Since we obtain a $p - value < 0.05$ in the F-test we have strong evidence to suggest that there is a statistically significant difference in variance of the top and bottom 20 NBA players in the 2014 season. Furthermore in the Welch t-test we have a $p - value < 0.01$, therefore we have strong evidence to suggest mean dribbles before a shot differs between groups. Hence we can conclude that taking slightly longer before a shot is beneficial to a players shot accuracy as seen in Figure 2.

3 Home advantage

In Section 3, we will explore the effect that the location of a game has on its result. Our initial hypothesis is that a home game is more likely to lead to a winning result, and we will look at several aspects of a basketball match to prove or disprove this. Section 3.1, investigates this by cross-referencing the `LOCATION` data with other variables provided in the data set. These include `WIN_LOSE`, `SUCCESS`, `PTS_TYPE` and others. We will perform a range of statistical tests for numerical analysis and create graphs for visual observation. The second section, 3.2, summarize the results from Section 3.1 using logistic regression to create a box plot to visualise how significant some of the variables that we previously look at in Section 3.1 are, on the result of a game. Section 3 uses data from 'NBA'. As stated in Section 1, this contains data on game activity from the 2014 to 2015 NBA season.

3.1 Exploring Home "Advantage" on several variables

3.1.1 Win Rate

```
## [1] "Overall home win percentage: 56.5124257114787"
```

This result is the home win percentage for all 30 teams in the data set. Over half of the games played at home were won so this may indicate some correlation however, we cannot be completely sure and will run further tests to examine this result.

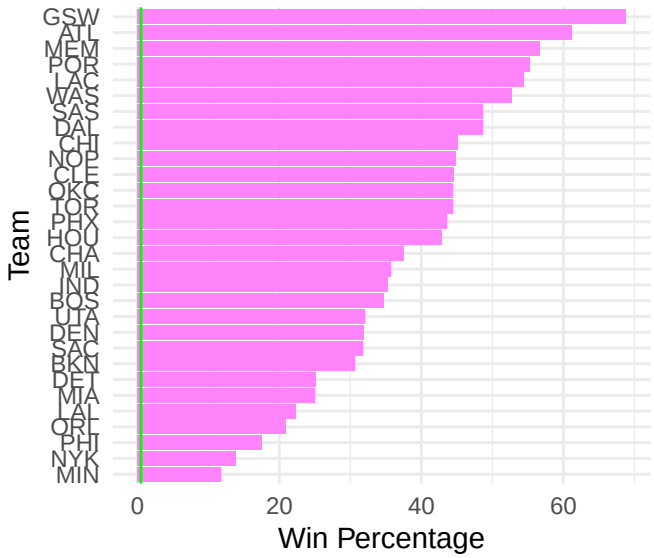

Figure 3: A bar chart showing the percentage of games that were won at home in the 2014/2015 season by team.

Figure 3 shows us that the Golden State Warriors have the highest home advantage, with a win percentage of close to 70%. However, the Minnesota Timberwolves are performed the worst when playing at home, with a win percentage of around 12%. Upon observation, we see no correlation between playing at home and winning. We will now perform a chi-square test to obtain a p-value to confirm this.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  win_loss_table
## X-squared = 1828.6, df = 1, p-value < 2.2e-16
```

Since we are comparing categorical variables, we use a Chi-Squared test to measure significance. A large X value of $X = 1828.6$ suggests a greater difference between observed and expected values. A p-value this small (essentially 0) suggests that the result is highly statistically significant, and may contradict the initial observation we make in Figure 3.

Therefore, we reject the null hypothesis. There is strong evidence of an association between playing at home and winning.

3.1.2 Shot Accuracy

This section looks at whether shot accuracy is affected by playing at home. We will use the `LOCATION` and `SUCCESS` variables for the overall effect, then use the variables `PLAYER_ID` and `PLAYER_NAME` to see how playing at home affects individual player's shot accuracy.

Values found from a shot accuracy analysis show that, in total, 186 fewer shots were attempted at home games. The accuracies of these shots look very similar at 0.45 for Away and 0.458 for Home. 'Accuracies' refers to shots that resulted in points being scored, or a `SUCCESS`. We will verify the statistical significance by doing a T-test, since we are comparing two means.

```
##
##  Welch Two Sample t-test
##
## data:  SUCCESS by LOCATION
## t = -2.7535, df = 124360, p-value = 0.005897
## alternative hypothesis: true difference in means between group A and group H is not
## 95 percent confidence interval:
##  -0.013308885 -0.002240557
## sample estimates:
## mean in group A mean in group H
##       0.4501325       0.4579072
```

We assume the null hypothesis, that the mean of A (Away) = mean of H (Home), is true. Although the means are very close in value, the t-test reveals a p-value of $0.005897 < 0.05$ so we reject the null hypothesis as there is evidence to show a significant difference in the average success rates between Group A and Group H. While the difference is $< 1\%$, the very large sample size makes this surprising result reliable.

A paired t-test, with the null hypothesis that the mean difference between home (H) and away (A) values is equal to zero, gave the mean difference to actually be 0.006664. This positive value means that players performed slightly better at home rather than away, on average. The confidence interval $(0.00015, 0.01318)$ does not include zero, which supports the significance of the result.

Table 2: Summary of the top ten players with the largest home court shooting advantage

Name	Away	Home	Diff	Team
Ed Davis	0.5176471	0.7328244	0.2151774	LAL
Nerlens Noel	0.3696498	0.5524862	0.1828364	PHI
Udonis Haslem	0.3600000	0.5254237	0.1654237	MIA
Bojan Bogdanovic	0.3312102	0.4842105	0.1530003	BKN
Mike Miller	0.2380952	0.3725490	0.1344538	CLE
Brian Roberts	0.3426966	0.4583333	0.1156367	CHA
Rasual Butler	0.3796791	0.4951456	0.1154665	WAS
Boris Diaw	0.4040816	0.5172414	0.1131597	SAS
Leandro Barbosa	0.4256757	0.5384615	0.1127859	GSW
KJ McDaniels	0.3443396	0.4554455	0.1111059	PHI

An interesting observation to be made from Table 2 is that of the top ten players with the biggest home court shooting advantage, only one was in the Golden State Warriors team, which we saw had the highest win rate at home in the 2014 to 2015 season in Figure 3. This leads us to conclude that player shot accuracy is not the most significant factor in the win or loss of a match. In fact, Nerlens Noel played for the Philadelphia 76ers, which was one of the three worst performing teams in Figure 3.

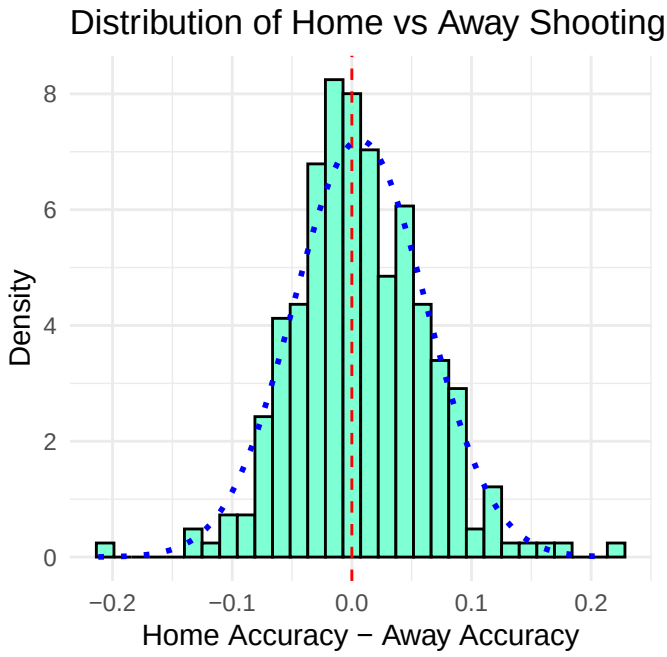

Figure 4: Histogram showing the distribution of differences in shooting accuracy between home and away games. A red dashed line indicates no difference, and the dotted curve shows a fitted normal distribution.

Figure 4 aligns well with the curve, suggesting the differences follow an approximately normal distribution. The vertical line represents no difference (i.e., home and away shooting accuracy are equal). The fact that the distribution is slightly skewed right of this line supports the idea of a home advantage, but a small one. Most values are within 0.1, but a few players show more extreme differences (both better and worse at home), indicating variability in individual performance, which is expected.

3.1.3 Shot Type

We will use the `PTS_TYPE` variable to analyse the shots taken inside or outside the arc, earning 2 or 3 points respectively.

```
##
##  Welch Two Sample t-test
##
## data:  PTS_TYPE by LOCATION
## t = -2.1858, df = 124350, p-value = 0.02884
## alternative hypothesis: true difference in means between group A and group H is not
## 95 percent confidence interval:
##  -0.0103270203 -0.0005624076
## sample estimates:
## mean in group A mean in group H
##        2.258370        2.263815
```

A shot-type analysis, calculated by dividing the successful shots by the overall number of shots, shows that the accuracy of 2-point shots at home is 0.494 compared to 0.487 away. The 3-point shots at home had an accuracy of 0.358 and 0.345 away. Both values are higher at the home location but the difference is small. This shows a similar outcome to Section 3.1.2. But the t-test showed a p-value of $0.02884 < 0.05$ along with a 95%confidence interval of $(-0.0103270203, -0.0005624076)$, both of which are statistically significant and allow us to reject the null hypothesis that the true difference in means between group A and group H is not equal to 0.

3.2 Logistic Regression

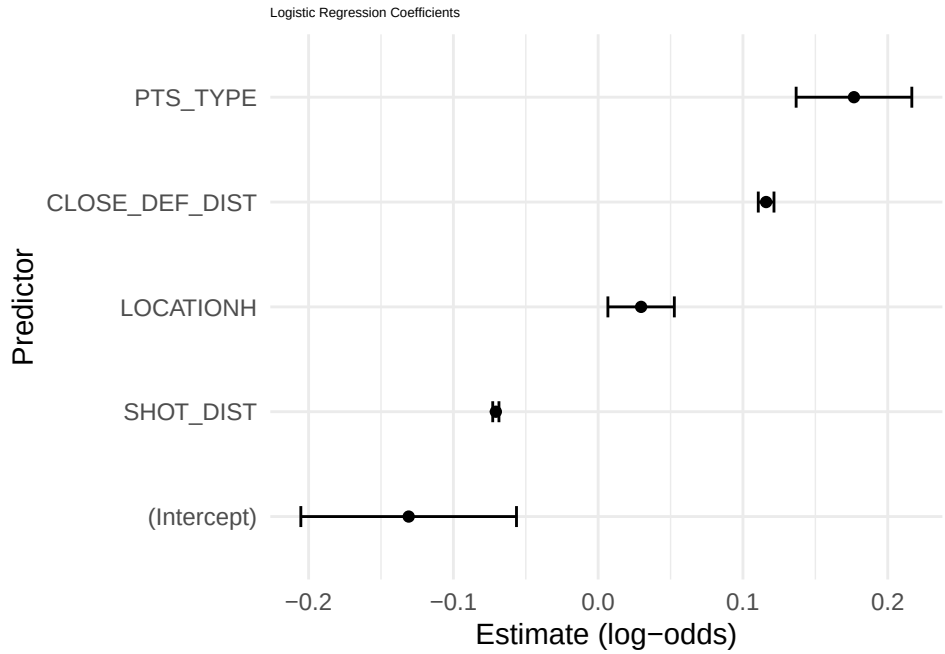

Figure 5: Estimated logistic regression coefficients (log-odds) with 95% confidence intervals for predicting shot success based on shot distance, defender proximity, shot type, and location. Positive coefficients indicate variables associated with a higher likelihood of a successful shot, while negative coefficients indicate a lower likelihood.

The logistic regression and box plot in Figure 5 allow us to make the following conclusions:

LOCATIONH (0.029661):

This is positive and statistically significant $p = 0.011255 < 0.05$. Playing at home increases the log odds of making a shot by 0.03 compared to away games. Converting to odds ratio: $exp(0.029661) \approx 1.03$, meaning players have about 3% higher odds of making shots when playing at home This supports the home advantage hypothesis, though the effect is relatively small.

SHOT_DIST (-0.070680):

This shows a significant negative effect $p < 2e - 16 \simeq 0$. Each additional unit of distance decreases the log odds of success by about 0.07. For each unit increase in shot distance, the odds of making the shot decrease by about $6.8\%(1 - exp(-0.070680))$. This confirms the idea that longer shots are harder to make.

CLOSE_DEF_DIST (0.115996):

We see this has a highly significant positive effect as $p < 2e - 16 \simeq 0$. As defender distance increases, shot success probability increases. For each unit increase in defender distance, the odds of making the shot increase by 12.3%. This shows that tighter defense significantly reduces shooting success.This makes intuitive sense as more open shots are more likely to go in.

PTS_TYPE (0.176674):

A positive effect $p < 2e - 16 \simeq 0$. 3-point shots have higher log odds of success than 2-point shots by about 0.18, and $exp(0.176674) \approx 1.19$, meaning 3-point shots have about 19% higher odds of success than 2-point shots when considering other factors. This might seem counter intuitive but makes sense when considering that SHOT_DIST is already in the model - this suggests that 3-point shooters may be more selective or skilled.

All variables show significance as their confidence intervals, indicated by the horizontal lines, do not cross 0. Additionally, their p-values are extremely small: approaching 0 and definitely below the 5% significance level.

4 Conclusion

Throughout this paper we have presented results of statistical tests from data on the $2014 - 2015$ NBA season.

- Section 2 This section explores variable relationships, Section 2.1 looks at Height and DRIBBLES and Section 2.2 looks at DRIBBLES and shooting_accuracy. With the use of ANOVA and Tukey in Section 2.1.1, we gained strong statistical evidence to show height has an impact of the number of dribbles a player takes before a shot, this could be attributed to position of player or playing style. For the analysis of DRIBBLES and shooting_accuracy multiple different tests were applied to the data obtained in Table 3 from this we are able to conclude that there is a relationship between the players with the best shot accuracy and tose who take longer dribbling before a shot.
- Section 3.1 Some may assume that a home game gives teams an advantage. This could be a result of knowing the court better, crowd support, or not needing to travel and therefore feeling more rested. Our hypothesis in Section 3 was tested in several ways and following this, we may conclude that a home court is not the primary factor in a win or a loss. The home advantage effect, while statistically significant in Section 3.1.1, is smaller than the effects of shot distance and the proximity of defenders. Teams should prioritise defensive pressure (reducing CLOSE_DEF_DIST), as seen in Section 3.2, as it has a stronger effect than home advantage.

5 Appendix

In Table 3 we compile the top and bottom 20 players in accordance to shot accuracy along with other contributing factors.

Table 3: Top and Bottom players in order of shooting accuracy.

PLAYER_NAME	total_attempted_shots	total_made_shots	shooting_accuracy	ave_dribbles	ave_defender_dist	Group
James Harden	1039	465	44.75457	4.9345525	3.760443	Top
Monta Ellis	1027	461	44.88802	3.2891918	4.286465	Top
Lamarcus Aldridge	1026	465	45.32164	1.2855750	4.060136	Top
Lebron James	970	477	49.17526	4.6886598	4.189278	Top
Russell Westbrook	962	420	43.65904	5.0478170	3.595634	Top
Klay Thompson	954	445	46.64570	1.6341719	4.131447	Top
Damian Lillard	953	410	43.02204	5.0293809	4.148059	Top
Stephen Curry	948	465	49.05063	3.7014768	4.489030	Top
Kyrie Irving	932	436	46.78112	5.1126609	3.639700	Top
Tyreke Evans	902	391	43.34812	4.5920177	3.533370	Top
Nikola Vucevic	896	475	53.01339	0.7243304	3.692746	Top
Blake Griffin	886	442	49.88713	1.1625282	4.618849	Top
Chris Paul	877	421	48.00456	6.0114025	4.574572	Top
Rudy Gay	870	391	44.94253	2.7011494	3.552069	Top
Gordon Hayward	857	385	44.92415	2.9708285	4.517153	Top
Kyle Lowry	853	354	41.50059	5.1465416	3.898124	Top
John Wall	843	379	44.95848	5.6714116	4.494187	Top
Anthony Davis	829	453	54.64415	0.5476478	3.741616	Top
Markieff Morris	817	383	46.87882	0.9241126	4.071848	Top
Pau Gasol	803	402	50.06227	0.9078456	3.765255	Top
Greg Smith	47	29	61.70213	0.1702128	2.691489	Bottom
Allen Crabbe	79	31	39.24051	0.4303797	5.697468	Bottom
Jerome Jordan	84	48	57.14286	0.3571429	2.688095	Bottom
Joey Dorsey	89	48	53.93258	0.4382022	2.621348	Bottom
Mike Miller	93	29	31.18280	0.3333333	6.291398	Bottom
Joe Harris	95	38	40.00000	0.6105263	5.611579	Bottom
Tyler Hansbrough	98	46	46.93878	0.2448980	2.812245	Bottom
Aaron Gordon	102	54	52.94118	1.3529412	4.127451	Bottom
Hedo Turkoglu	102	45	44.11765	1.0196078	5.242157	Bottom
Bismack Biyombo	114	64	56.14035	0.3157895	2.347368	Bottom
Robbie Hummel	114	56	49.12281	0.4649123	5.307895	Bottom
Jason Maxiell	128	54	42.18750	0.3984375	3.357813	Bottom
Alonzo Gee	129	63	48.83721	1.4341085	4.068217	Bottom
Udonis Haslem	134	58	43.28358	0.2164179	3.679105	Bottom
Chris Andersen	135	79	58.51852	0.1407407	3.380000	Bottom
Jimmer Fredette	147	56	38.09524	3.2517007	4.596599	Bottom
Luke Babbitt	153	72	47.05882	0.5686275	5.905229	Bottom
James Ennis	154	61	39.61039	0.9025974	4.527273	Bottom
Garrett Temple	157	62	39.49045	1.1401274	5.361147	Bottom
Jordan Farmar	157	60	38.21656	3.5095541	5.243949	Bottom