A study of the nba dataset for MAT022, School of Mathematics, Cardiff University

Subhro Mitra (Student ID: C1893753)

23 February 2021

Abstract

We demonstrate how various descriptive and inferential statistical methods and analysis can be applied to the National Basketball Association (`nba`), `2014-2015` dataset, and how their results might be interpreted and presented. Here we analysed the success of each player, the basketball teams and their respective performance. We showed the top fifteen best performances by graph in our player performance section. We illustrated the nba dataset and produced some insight into the best-performing player and best-performing ground. We evaluated closest defender, shot distance, dribbles, number of two or three pointers and shot distance by the players which produced the statistical evidence for performance of each player. We also performed inferential analysis, t-test and finally present the results of a logistic-regression analysis and demonstrated that to a certain extent it was possible to distinguish between the players by the scores and achieved a subset of the nba events.

Contents

1 Introduction

The National Basketball Association (NBA) is a North American professional basketball league with 30 teams. It is one of the four major professional sports league of United States and Canada. The given `nba` data set records the performances of teams in a span of 120 days for the year 2014-2015. This is a combined event of 'players 'home ground' and 'away ground'. According to the provided data, Los Angeles Clippers (LAC) with 4,855 total points was at the top. Stephen Curry had the record for the highest points (1,130) with over 570 points scored only with three pointer. The current nba world-record holder is the Philadelphia 76ers, which is an American professional basketball team based in the Philadelphia metropolitan area. For reference, following are the links for more information on nba: [https://en.wikipedia.org/wiki/Philadelphia_76ers] and [https://www.nba.com/standings]

The provided dataset consists of 1,28,069 observations of 24 variables. These names of the variables are listed hereafter:

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

```
##  [1] "GAME_ID"            "DATE"                "HOME_TEAM"
##  [4] "AWAY_TEAM"          "PLAYER_NAME"         "PLAYER_ID"
##  [7] "LOCATION"           "W"                   "FINAL_MARGIN"
## [10] "SHOT_NUMBER"        "PERIOD"              "GAME_CLOCK"
## [13] "SHOT_CLOCK"         "DRIBBLES"            "TOUCH_TIME"
## [16] "SHOT_DIST"          "PTS_TYPE"            "SHOT_RESULT"
## [19] "CLOSEST_DEFENDER"   "CLOSEST_DEFENDER_ID" "CLOSE_DEF_DIST"
## [22] "FGM"                "PTS"
```

2 Background

The NBA started life as the Basketball Association of America in 1946 and played under that moniker for 3 years before merging with the National Basketball League and changing names to the NBA in 1949. This game has many rules and regulations to make this hard to play. The original rule was published by James Naismith on 15th January, 1892, which is a bit different from the current rules as here was no dribbling, dunking, three-pointers, or shot clock, and goal tending was legal. The original rules for nba are given as follows:

1) The ball can be thrown in any direction. It can be done using one hand or both hands.

2) The ball can be struck in any direction. This also allows to use one or both hands.

3) The side with most of the points is declared the winner of the game.

4) A player cannot run with the ball. Player must throw it from the spot he/she catches the ball from. There is only an allowance for anyone catching the ball as while running.

5) The body must not be used to catch the ball. Player must catch the ball in between the hands.

6) A foul occurs when a player strike the ball with the fist. Also, holding, tripping, pushing or shouldering an opponent result into a foul.

7) There is no tripping, shouldering, striking, holding or pushing in any way of an opponent.

8) A goal is made when the ball is thrown or batted from the grounds into the basket. If the ball stays at the basket without dropping, that too is considered as a goal.

9) The referee will be the judge of the game. He keeps account of the baskets. Other than that, he/she decides when the ball is in play, which side it belongs and keeps time.

10) The time for any basketball game will be two fifteen-minute halves with a five minutes break in between.

11) If one or the other side makes three sequential entangles, it shall count a goal for their opponents. The most recent international rules of nba were approved on 2nd February, 2014 by FIBA and became effective 1st October, 2014. There are eight rules encompassing 50 articles, covering equipment and facilities, regulations regarding teams, players, captains and coaches, playing regulations, violations, fouls and their penalties, special situations, and the officials and table officials. The rules also cover officials' signals, the scoresheet, protest procedure, classification of teams and television timeouts.

source: https://en.wikipedia.org/wiki/Rules_of_basketball, https://sportsierra.com/nba-official-basketball-rules-and-regulations/ and https://nbahoopsonline.com/History/

```
##  [1] "GAME_ID"            "DATE"                "HOME_TEAM"
##  [4] "AWAY_TEAM"          "PLAYER_NAME"         "PLAYER_ID"
##  [7] "LOCATION"           "W"                   "FINAL_MARGIN"
## [10] "SHOT_NUMBER"        "PERIOD"              "GAME_CLOCK"
## [13] "SHOT_CLOCK"         "DRIBBLES"            "TOUCH_TIME"
## [16] "SHOT_DIST"          "PTS_TYPE"            "SHOT_RESULT"
## [19] "CLOSEST_DEFENDER"   "CLOSEST_DEFENDER_ID" "CLOSE_DEF_DIST"
## [22] "FGM"                "PTS"

## [1] 904

## [1] 120

## [1] 281

## [1] 30
```

Remark An entry of the data set consists of the home teams and away teams, with the number of Games904 which were played accross 120 days. There were 281 players played in 30 teams. In the given dataset, it is clear that the player 'Stephen Curry' scored maximum number.

3 Data Exploration and Inferential Analysis

3.1 Cleaning of Data

After studying the data, we found that we have touch time in negative and some discrepancies in data when the distance was greater than 23.75. So it can't be considered as a 2-pointer, it must be considered as a 3-pointer. Thus we cleaned our data accordingly.

Assumption Since the aggregate value of mismatch (10,336) contributes to *8%* of the dataset, therefore we dropped all NA and negative data and accordingly rectified the two pointer and three pointer values.

4 Descriptive Analysis

```
## `summarise()` has grouped output by 'PLAYER_ID'. You can override using the `.gr
```

PLAYER_ID	PLAYER_NAME	TOT_PTS_SCORED	TOT_THREE_PTS	TOT_TWO_PTS	TOUCH_TIME_M
201939	Stephen Curry	1113	555	558	3.83
202691	Klay Thompson	1057	519	538	2.33
201935	James Harden	1046	432	614	5.64
2544	Lebron James	1018	246	772	5.67
202681	Kyrie Irving	986	354	632	5.18
101145	Mnta Ellis	981	201	780	3.87

```
## [1] 281
```

Remark Here we have seen the number of scores and can conclude that the player 'Stephen Curry' scored maximum number of two points and three points.

4.1 Player Performance

In this section, we investigate the performance of each player.

4.2 Home Team Analysis

4.2.1 Home Team Performance

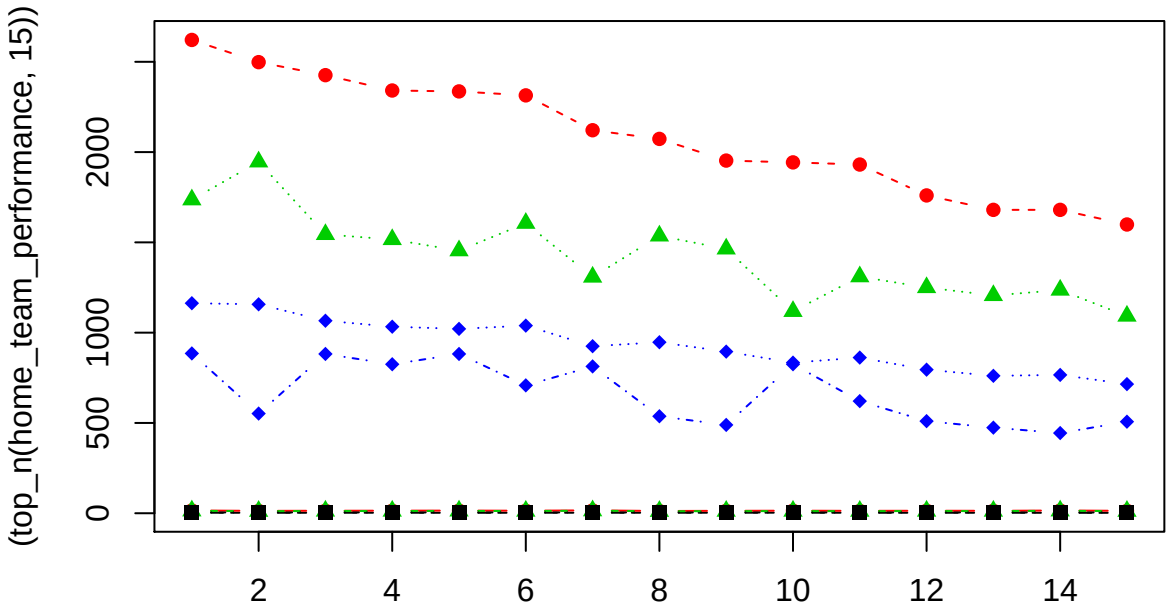
HOME_TEAM	total_points_scored	total_two_pointers	total_three_pointers	touch_time_mean	shot_distance_m
LAC	2621	1736	885	2.754217	15.33
WAS	2498	1946	552	2.631928	12.96
ATL	2426	1544	882	2.670872	13.82
SAC	2360	1910	450	2.886445	12.50
PHX	2341	1516	825	2.873483	14.06

```
## [1] 30
```

```
## Selecting by close_defender_distance_mean
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log = log): NAs introduced by
## coercion
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion
```



4.2.2 Home Team Wins

```
## `summarise()` has grouped output by 'GAME_ID'. You can override using the `.grou
```

GAME_ID	HOME_TEAM	total_points	touch_time_mean	total_wins
21400741	GSW	107	2.365432	81
21400302	WAS	103	2.567442	86
21400550	LAC	103	2.535714	84
21400636	GSW	103	2.414286	84
21400382	GSW	102	2.313793	87

```
## [1] 503
```

Remark GSW has gained the most number of points in home ground.

4.2.3 Home Team Loss

```
## `summarise()` has grouped output by 'GAME_ID'. You can override using the `.grou
```

GAME_ID	HOME_TEAM	total_points	touch_time_mean	total_wins
21400234	LAL	102	3.068235	85
21400164	BKN	97	2.394318	88
21400565	TOR	97	2.256626	83
21400125	BOS	96	1.663636	77
21400699	BKN	96	2.923913	92

```
## [1] 393
```

Remark LAL is the team who lost most number of times in home grounds.

4.3 Away Team Analysis

4.3.1 Away Team Performance

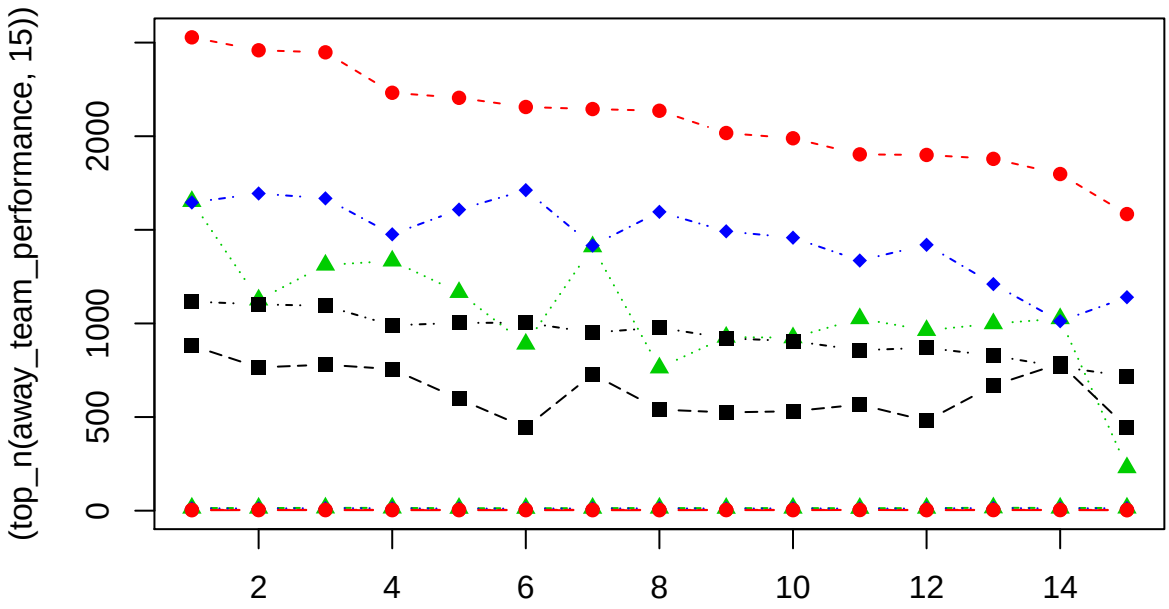
AWAY_TEAM	total_points_scored	total_wins	total_two_pointers	total_three_pointers	touch_time_mean	sho
GSW	2528	1652	1646	882	2.440150	
PHX	2459	1126	1694	765	2.959659	
LAC	2448	1312	1668	780	2.861378	
BKN	2236	961	1684	552	3.124118	
DAL	2232	1334	1476	756	2.668232	

```
## [1] 30
```

```
## Selecting by close_defender_distance_mean
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log = log): NAs introduced by
## coercion
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion
```



4.3.2 Away Team Wins

```
## `summarise()` has grouped output by 'GAME_ID'. You can override using the `.grou
```

GAME_ID	AWAY_TEAM	total_points	touch_time_mean	total_wins
21400257	DAL	102	2.873404	94
21400271	TOR	102	4.396154	78
21400442	PHX	101	3.260000	90
21400198	LAC	100	2.913253	83
21400145	GSW	99	2.352000	75

```
## [1] 393
```

Remark DAL is the away team who scored the highest points.

4.3.3 Away Team Loss

```
## `summarise()` has grouped output by 'GAME_ID'. You can override using the `.grou
```

GAME_ID	AWAY_TEAM	total_points	touch_time_mean	total_wins
21400302	BOS	100	2.867033	91
21400749	GSW	100	2.073684	95
21400480	PHX	96	4.105063	79
21400046	BOS	95	2.203371	89
21400248	TOR	95	3.478788	99

```
## [1] 503
```

Remark BOS is the team who has won the most number of time in away grounds.

5 Overall Remark

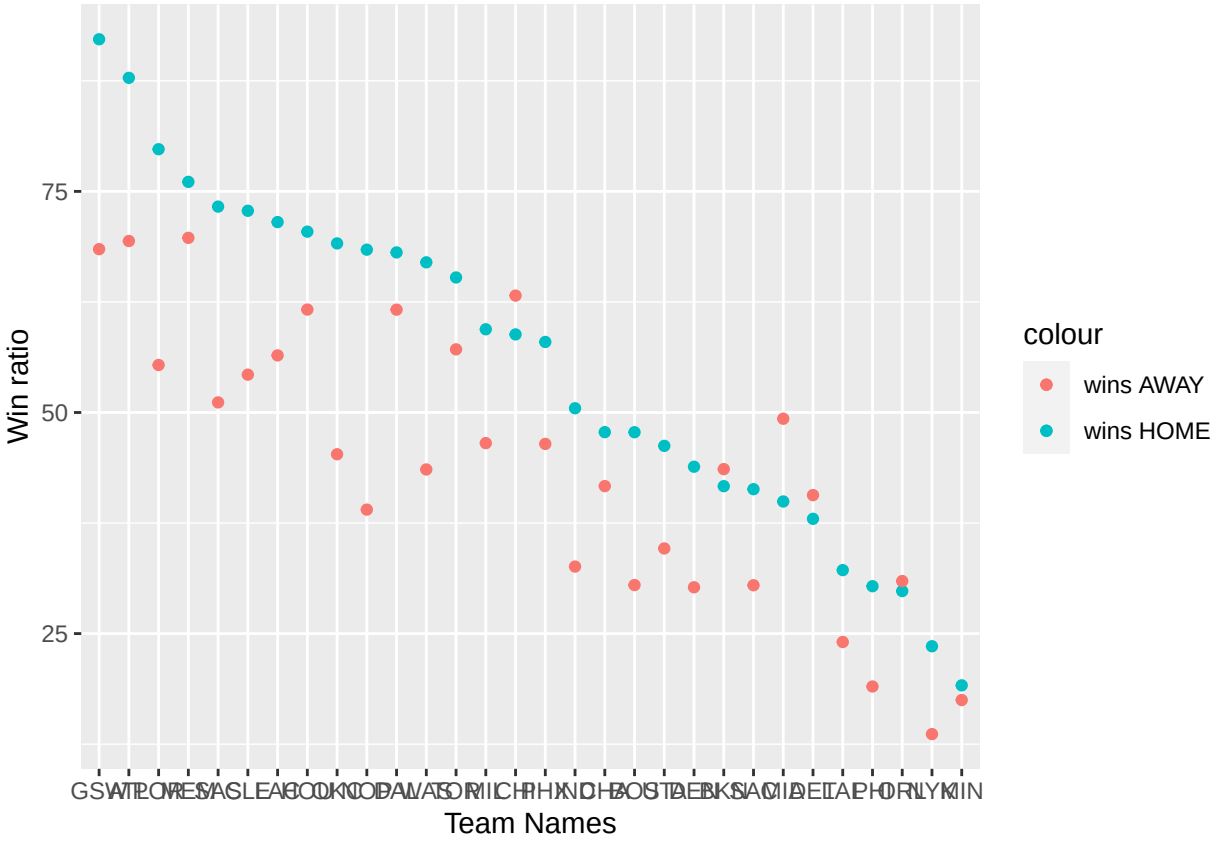
The teams playing in Home Ground has more wins than the teams playing in Away Ground. LAC has the highest overall wins followed by ATL and SAC. Stephen Curry scored the most number of points followed by Klay Thompson and James Harden

6 Correlation

In our analysis, we must pay special attention to the total points scored, total two and three pointers, shot distance and closest distance. Here shot distance and touch time played a significant role in player's performance.

```
##
##  Welch Two Sample t-test
##
## data:  df_team_perf$total_away_wins and df_team_perf$total_home_wins
## t = -2.2994, df = 55.25, p-value = 0.02529
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -456.45013  -31.34987
## sample estimates:
## mean of x mean of y
##  954.6333 1198.5333
```

6.1 Impact of Touch Time on Shot Success Percentage



Remark After performing ggplot, we observed that success percentage is highly correlated to touch time. We found the best time for shot to be success is in the range of 3.8 and 4.7.

7 Conclusion

We have done one hypothesis testing which produces an outcome that our hypothesis was wrong as the number wins depend on grounds, i.e., number of wins are greater in home grounds, whereas it is less in away grounds. According to the provided data, defender distance is significant to calculate shooter's success probability. Touch time and shot distance played a vital role for team's performance. So analyzing team's weakness and filling the gap would be easier.

7.1 Home Vs Away Wins

We calculated the number of wins, number of losses and percentage of wins for both home teams and away teams. After that stuff, we merged these two data in one dataframe on the basis of team. Later we arraged home win percentage data in decreasing order.In the dataframe 'df_team_per', we added a new column named 'WIN_PCT_DIFF', which calculated the difference between home win and away win of each team. Now we are applying t-test for total home wins and total away wins for individual teams as our hypothesis was that home ground and away ground has no effect on the winning, which is proved wrong after the t-test.

Now we created a 'ggplot' which clearly shows that home wins are at the up and away wins are at the down. This means the number of home wins are greater than the number of away wins. Thus our null hypothesis got rejected and the testing is done.

8 Markdown Test

8.1 Hyperlinks

1) https://en.wikipedia.org/wiki/Philadelphia_76ers.

2) https://www.nba.com/standings.

3) https://en.wikipedia.org/wiki/Rules_of_basketball.

4) https://sportsierra.com/nba-official-basketball-rules-and-regulations/.

5) https://nbahoopsonline.com/History/.

6) https://github.com/rstudio/rmarkdown.

7) https://bookdown.org/yihui/rmarkdown/.

8) https://rmarkdown.rstudio.com/articles.html.

9) http://libgen.rs/book/index.php?md5=4A250C4FEC4D02287017AD3B71EAA34F.

8.2 Citations

See for example (Grimmett and Stirzaker, 2001) and (Hogg et al., 2005) and (Stern, 2010) and (Levin and Rubin, 1994)

References

Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and Random Processes*. Oxford University Press, third edition.

Hogg, R. V., McKean, J. W., and Craig, A. T. (2005). *Introduction to Mathematical Statistics*. Prentice Hall, sixth edition.

Levin, R. I. and Rubin, D. S. (1994). *Statistics for Management*. Prentice Hall, sixth edition.

Stern, D. (2010). *Official Rules of the National Basketball Association*. Prentice Hall.