A study of the NBA dataset for season 2014-2015

Stefanidou Evdoxia

23 February 2021

Abstract

We demonstrate how various descriptive and inferential statistical methods can be applied to the NBA dataset for the season 2014-2015 to get useful insights from the data. We study two main factors that can affect the outcome of a game, the home-court advantage and the rest days between the games. Both of them show that they significantly affecting the outcome of the game for the home team but do not indicate the same for the away team. Also, we give some insight for the two factors at the teams level.

Contents

1 Introduction

National Basketball Association is a men's professional basketball league based in North America. The league involves 30 competing teams, playing throughout October to April for the regular season, and each team is playing 82 games. This report's primary purpose is to analyze the

data from 2014–15 for the regular season of the National Basketball Association (NBA), a total of 904 games. The data contains information about 30 teams, 128 thousand NBA shots (rows) and 23 different characteristics considering the shots (columns). The name of the columns are self-explanatory and are presenting below :

```
## [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"
```

In this study, we will focus on "W", "LOCATION", "GAME_ID", "HOME_TEAM", "AWAY_TEAM", and "FINAL_MARGIN" and we will try to analyze them to get insights concerning the game outcome. Driven from the dataset's information, we can produce meaningful results about whether some factors are crucial for winning an NBA game and how much advantage they provide in the final game outcome.

2 Background and Purpose of the Study

Our purpose in the first phase is to investigate whether playing at "home" or "away" is a factor that affects the outcome of a game in the session of 2014-2015. Based on the literature review, home-court advantage is mainly affected by the game location, travel status, psychological and behavioural factors (Carron et al.,2005). When a team plays at home court is estimated to have a 4.68 ± 0.28 points advantage against the "away" team (Harville and Smith,1994). This advantage is mainly credited to the crowd's support and the confidence of winning the home players (Schwartz and Barsky, 1977). We aim to investigate if this phenomenon appears in our dataset by focusing on the "location" of the teams. Therefore, drawing upon the literature findings, we argue that the means on winning ratios of "home" and "away" win does not show any significant difference, H0:mA=mB where ma= mean of winning ratio in home court and mb= mean of winning ratio in away court(Null Hypothesis). We aim to estimate the amount of home-court advantage, if it exists in this dataset.

In the second phase, we want to explore if rest days affects the outcome of a game. Based on the literature review, one main factor affecting the outcome and the home-court advantage is the travel factor and specifically how number of rest days affect the game's outcome (Courneya and Carron, 1992). Home teams have an advantage over the away teams. The NBA schedule ensures that home teams play on average once every two days. Away teams have a more tight schedule as they play back to back games with only one day gap between the games used for travelling to the opponent team's home court (Nevill and Holder, 1999). As a consequence of this schedule, visiting or away teams are more tired than their home team opponents. Therefore, we will try to identify if rest days can explain a part of a game's outcome .

3 First part: Effect of home court advantage on game outcome

3.1 Descriptive analysis and methodology

The first part of the analysis focuses on investigating whether playing at home court gives an advantage at the overall outcome of the game and at team level. Also, we will try to identify how much is the actual home-court advantage. Therefore, we transformed the original data to extract the necessary informations. We encode the column "W" from the original dataset, to 1 for win and 0 for losing . Then we create a new column named "WINNER", which have as values "HOME" when the home team wins and "AWAY" when the away team wins. A new data frame is created containing 6 features ("GAME_ID","HOME_TEAM","AWAY_TEAM","LOCATION", "FINAL_MARGIN","WINNER"). We strip out duplicates by "GAME_ID" and the dataframe does not contain any missing values. We further analyze this data to extract useful informations and finally create a data frame containing "TEAMS", "HOME_WINS", "AWAY_WINS", "TOTAL_HOME_GAMES", "TOTAL_AWAY_GAMES", "HOME_WIN_RATIO", "AWAY_WIN_RATIO", "DIFFRENCES_%" and "TOTAL_WIN_RATIO". The name of the columns is self-explanatory. The ratios were calculated for each team as the number of home/away wins expressed as a percentage of the total number of games played at home/away. The "DIFFRENCES_%" column contains the difference of home and away win ratios, and finally, the "TOTAL_WIN_RATIO" is calculated as the total number of wins expressed as a percentage of the total games played by each team. The dataset has 30 rows corresponding to the teams. Finally, we set the significance level at 0.05, which means that if the chosen statistic resulted in a p-value higher than the alpha level, we do not have enough evidence to reject the null hypothesis.

Table 1: The number of games won by away and home teams and their respective percentages.

row_labels	Count	Percents.
WINS BY :|AWAY TEAM	398	44.02655
WINS BY :|HOME TEAM	506	55.97345
WINS BY :|#Total cases	904	904.00000

An inspection between the number of home and away wins as the Table 1 shows is 56% of the games were won by home teams which account for 108 more games from the away wins. The median of home final margin is 2 implying that there might be is a home-court advantage. Also, from the inspection of each team, as Figure 1 shows, 28 teams had more wins in the home court with a maximum difference of 10 from the "POR" team. Only two teams had one more win in away court in respect of their home games. These factors indicate evidence that home-court advantage might be present in this dataset and that the mean home and away win ratios are a sufficient variables for identifying it. Also, the number of games played by each team in home and away courts are not equal. Therefore, the winning ratio will also allow us to compare any existing difference between them.

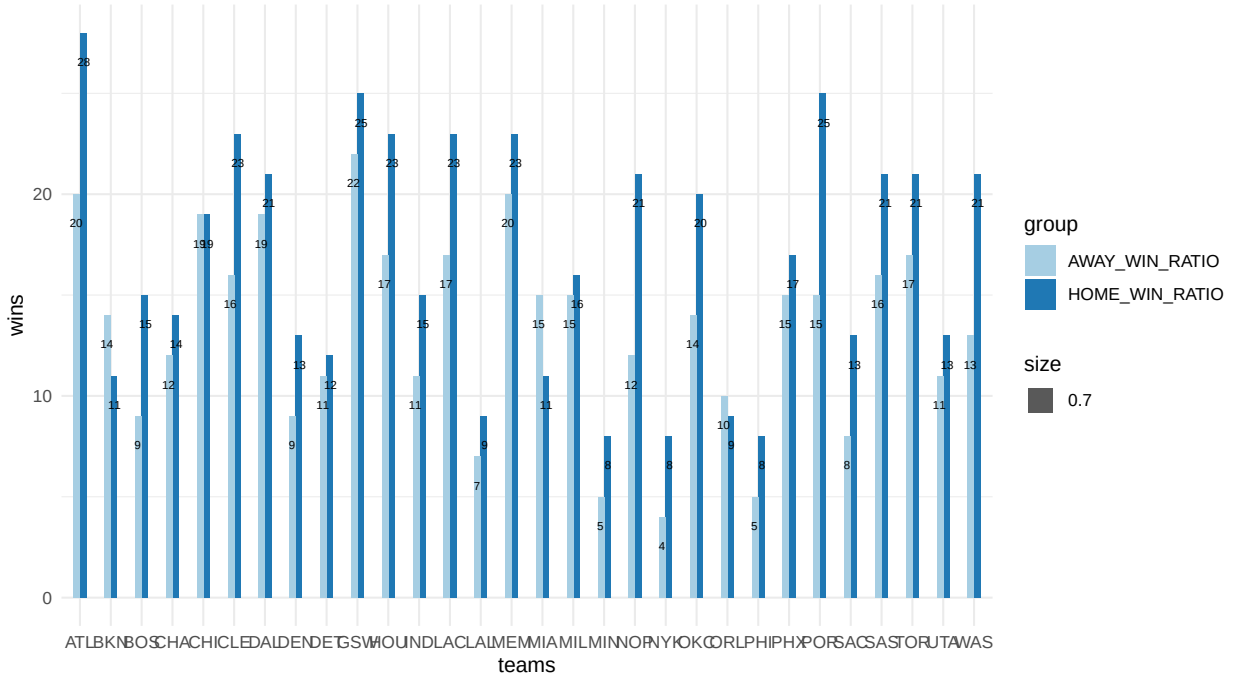

Figure 1: Bar Plot showing wins of each team in home and away court (#fig:ggplot)

3.2 Examination of in respect of the game outcome

We want to examine whether playing at home court give an advantage at the outcome of the game. Therefore we will compair the the means for home and away wins with the independent Student's t-test, test to to evaluate whether the means of home win ratio and away win ratio does not show any significant difference with H0:mA=mB, and alternative hypothesis Ha:mA>mB (greater) that the means of home win ratios is significantly different and greater than the means of away win ratios. The Shapiro-Wilk normality test is applied in both subsets with the two p-values=0.2022 and p-values= 0.7004 both to be greater than the significance level 0.05 indicating that the distribution of the subsets are not significantly different from the normal distribution as we do not have enough evidence to reject H0. Therefore, we can assume the normality. F-test to test for homogeneity in variances is applied between the subsets with p-value of p = 0.3078. It's greater than the significance level alpha = 0.05 implaying there is no significant difference between the variances of the two sets of data. Therefore, we can use the classic t-test which assume normaly distributed subsets and equality of the two variances. The value of the t-test is t = 2.6022 with degrees of freedom df= 58 , p-value= 0.00587 and confidence interval of the means at 95%. P-value is less than the significance level alpha = 0.05. We can conclude that the mean of home win ratios is significantly different and higher than the means of the away win ratios. Therefore the home court advantage existis in this dataset and influence the game outcome. The Figure 2 clear illustrates what t-test found.

The performance of each team is found and defined as the total wins a team had expresed as a persentage of total games played. In this way we can order the teams by their ability and find the best and worse teams of the regular season. The best eight teams are "CLE", "DAL", "LAC", "HOU", "POR", "MEM", "GSW", "ATL" and the eight with the lower performance are "DET",

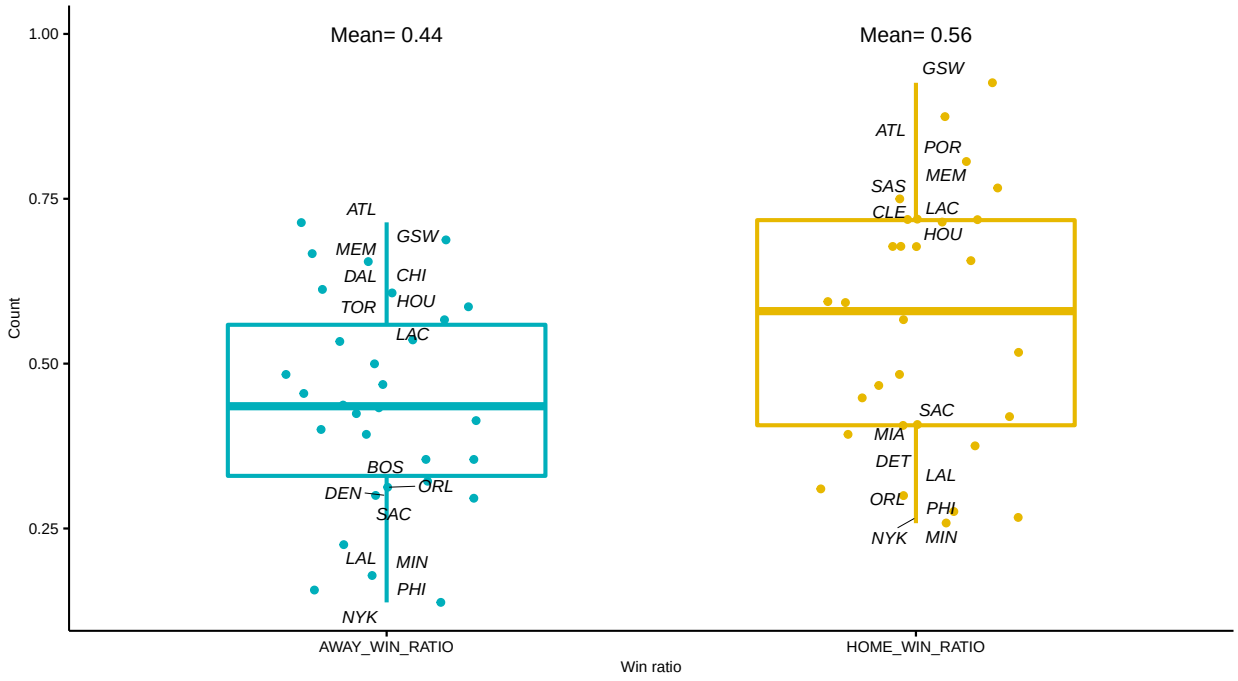

Figure 2: (#fig:plotofqq)Boxplot to visualize the means of the home and away win with their respective observations appearing on the plot and the outliers have their teams labels visible.

"DEN", "SAC", "ORL", "LAL", "MIN", "PHI", "NYK". Compairing them with the outliers is obvious that in both cases the upper outliers refering to best performed teams and the lower outliers to the lower performed teams. For thoses cases we assume that home court advantage is not signifintant as the best teams can perform equaly well on home and away court without the need of home court advantage and in the other case that those teams do not have many wins regurdless the home court advantage. This assumption can be tested by analising the difference between the home win ratio and the away win ratio of each team as Table 2 shows. We can clearly see that the differece between the ratios of home and away teams for the 25 teams out of the 30 is positive, indicating that they have more wins when playing at home confiring the t-test outcome. Only 5 teams have won more games in the away courts with maximun of 9.1% difference. There is a flacuation in the differences among the teams but there is a clear upturning trend from lower to higher performace teams indicating that home court advange is stronger in the best performing teams. If we seperate the teams into two categories best performnce (15 teams) and lower performace(15 teams), we can see that for the best performance teams the higest difference of 29% is appearing in "OKC" team which is in posision 12 and for the lower performance teams the highest difference is 16,2% appearing in "BOS" team which is in posision 16 of the overall rank. This indicates that the home advantage is stroger in best permorfance teams.

Table 2: Summary table. The first 5 rows represent tha data explored for the 5 best performed teams and the rest rows showing the data of the 6 teams with lower performance .

	TEAMS	HOME_WIN_RATIO	DIFFRENCES_%	TOTAL_WIN_RATIO	ADVANTAGE_adj
8	ATL	0.8750000	16.071429	0.8000000	0.8850920
13	GSW	0.9259259	23.842593	0.7966102	0.8814129
17	MEM	0.7666667	10.000000	0.7166667	0.7946470
20	POR	0.8064516	27.073733	0.6779661	0.7526437
3	HOU	0.7187500	11.160714	0.6666667	0.7403800
28	SAC	0.4062500	10.995370	0.3559322	0.4031275
1	ORL	0.3000000	-1.250000	0.3064516	0.3494242
12	LAL	0.3103448	8.453838	0.2666667	0.3062440
9	MIN	0.2580645	7.949309	0.2203390	0.2559627
5	PHI	0.2758621	11.961207	0.2131148	0.2481220
15	NYK	0.2666667	12.873563	0.2033898	0.2375671

As the above analysis outcome shows, the home-court advantage affects the game's outcome and shows higher values on the best-performed teams. To justify this outcome, we need to consider the ability (performance= total win ratio) of each team. A team's ability can outweigh the home advantage of its opponent if this team's ability is significantly lower. Therefore we want to find the adjusted home-court advantage to be able to compare its effect among the teams. We consider this value as the actual home court advantage. The home-court advange is calculated as the home win ratio. Therefore to find the home advantage for each team, we conduct a linear regression with the dependent variable of the home win ratio and as an independent variable the ability (performance) of each team. The coefficient (b) of the team' ability can be interpreted as the average effect on the home win ratio. Therefore, the regression equation without the residuals can predict the home advantage adjusted to the team's ability. The regression model's standardised residuals show if the team's home advantage is above or below the home advantage predicted by the regression equation based on that team's ability. The summary indicates that the ability of a team is significantly associated with the dependent variable. The model is statistically good and reliable as Shapiro-Wilk test for normality, Durbin-Watson test for autocorrelation, and Breusch-Pagan test for homoscedasticity for the residuals have p-value=0.7062, p-value= 0.3028 and p-value= 0.4273 respectively, suggests that there is no evidence that the residuals are not normally distributed, uncorrelated and heteroscedastic. Also, the model is well fitted to the data. The predictions made using the test data to evaluate the performance show Root Mean Squared Error = 0.0465 and R-squared and adjusted R2 = 0.95.

Regression Equation:

HOME_WIN_RATIO = 0.01682 + 1.08534∗TOTAL_WIN_RATIO

The coefficient (b) of team' ability can be interpreted as the average effect on the home win ratio. Therefore, an increase in the ability of 10 percentage points would suggest an increase in the home advantage of almost ten percentage points.

After adjusting the home advantage for each team, there are significant differences between the adjusted and not adjusted home advantage. This indicates that the first observed value is influenced by the team's ability and do not present the actual values. Sixteen teams show a lower adjusted

advantage than the original observation, and the rest of them a positive one. No team is observed to have the same adjusted and not home-court advantage indicating that the adjusted value is significant influence the home advantage and, therefore, the outcome of the game. The previous upturning trent from the lower to the better performed also exists, giving us evidence to support that the home advantage has a higher contribution on the best performing teams. Also, the adjusted home advantage shows a gradual reduction from best to lower performed teams without any overlap as they exist in the first observed value. Moreover, the bottom two performance teams showed a significant reduction of almost 3.3%. The biggest increase of 10% was observed in "CHI", which is also the only team with equal wins in home and away courts.

4 Second part: Effect of rest days on game outcome

4.1 Descriptive analysis and methodology

The second part of the analysis focuses on investigating whether the number of rest days between games affects the game's final outcome and, specifically, the team's winning ratio. A dataset is created containing the `GAME_ID`, `DATE`, `HOME_TEAM`, `TAWAY_TEAM` and `WINNER` columns without containing missing values. The date was transformed in order to allow us to find the number of rest days defined as the difference between the date of the current match and the date of the previous match minus one. We subtract one day as we consider it the travelling date. The rest days found are 0, 1, 2, 3, 4, 7, 8, 9, 10 without containing the fist 30 games played which were given a discreet value of 30 in order to do not interfere with the real rest days. As a result, we have 1805 observations of the game played. The formula applied in ratios used in this part is the number of corresponding wins played after rest days (total wins, home wins, away wins) expressed as a percentage of the total number of games played after number of rest days, where is the respective number of rest days found. Finally, we set the significance level at 0.05, which means that if the chosen statistic resulted in a p-value higher than the alpha level, we do not have enough evidence to reject the null hypothesis.

An inspection between the rest days of home and away teams among their games shows that away teams play more than twice back to back games (300) than the home teams (127), which can indicate that rest days might have a significant contribution to home advantage and game outcome. On average, teams have 0 or 1 day of rest between their games as the rest days distribution indicates shown in Table 3. Home teams play about 122 games with one more day of rest and 28 games with two more days of rest than their opponents as shown in Table 3.

Table 3: Frequency of appearance of home and away teams in respect of the 11 categories of rest days.

	0	1	2	3	4	5	7	8	9	10	30
AWAY	300	434	113	23	4	2	4	7	3	1	13
HOME	127	556	141	36	9	1	2	13	0	2	17

4.2 Examination of rest days in respect of the game outcome

Analysis is conducted based on the total, home and away wins. The overall winning ratio in respect of rest days shows small variation for the 0,1,2,3 rest days with values approximately at 50%. For rest days between 4 and 8 winning ratio shows the highest values. Above eight days, the ratio reveals unbalanced fluctuation, and we consider that the effect of rest days do not contribute to the game outcome ,but the outcome is a result of the team's ability. An approximate same inverse behaviour is observed form the overall loss ratio. This indicates that the number of rest days when between 0 and 8 can affect a game's outcome. Specifically, the distribution of 0,1,2,3,4 rest days is expressing 95% of the total observation as Table 4 shows. Therefore we conclude that resting days between 0 and 4 can have a higher impact on the outcome of a game and the further analysis is based on them.

We are interested in analyzing the distribution of the wins concerning the rest days and in examining if they are statistically significantly associated. The Pearson's chi-squared test is selected, as our data referring to categorical variables (number of rest days, the game outcome) with Null hypothesis(H0): number of rest days between the games and number of wins for each rest day category are independent. The Pearson's Chi-Square test has X2 (4) = 6.9637 and p= 0.1378 higher than the level of significance alpha = 0.05. Therefore, we conclude that the rest days and corresponding wins are independent as we do not have enough evidence to reject the H0.

Table 4: Distribution of each category of the rest days in respect of home and away teams.

	Rest_Days	Distribution_Of_Rest_Days	Win_Ratio	Home_Win_Ratio	Away_Win_Ratio
4	0	23.6565097	0.4707260	55.03513	44.96487
3	1	54.8476454	0.5010101	55.15152	44.84848
5	2	14.0720222	0.5393701	56.69291	43.30709
2	3	3.2686981	0.4745763	55.93220	44.06780
7	4	0.7202216	0.7692308	76.92308	23.07692

Based on the indications of the descriptive analysis that a potential relation exists we continue the analysis. As the Table 4 shows, overall wining ratio shows a gradual increase from 47.07% to 50.10% to 53.93% and 76.92% respectively for 0, 1, 2 and 4 days of rest, indicating positive relation. Win home ratio has a similar upturn. The away win ratio has a different behaviour showing hardly any difference between 0,1,2,3 days of rest with approximately 44.5% but the fourth rest days shows a significant drop in away wins. The reason for this decline can be due to each team's ability, the lack of court advantage or because those rest days are appearing in the middle of the season and teams are tired. A further study upon these factors is suggested to understand the away win ratio further.

The balloon plot of Figure 3 shows the high relative magnitude that 0,1,2 and 3 rest days have in relation to the outcome of a game as explained above and provide us with more evidence that those rest days are the corresponding main component of the rest days distribution. Therefore, we conclude that we have some evidence to indicate that the rest days and the game outcome show some dependence between them when the examination is done regarding the home wins.

Figure 3: A graphical matrix of the distribution of games outcomes (loss or win) based on the rest days categories. Each cell contains a dot whose size reflects the relative magnitude of the corresponding component and the color is based on the obsevations each component holding (#fig:balloonplot)

5 Conclusion

In this report, we examine two main issues. First, we examine whether playing at home court gives an advantage in the game's outcome, and it was concluded that a team playing at home is more likely to win with an average win probability of 56%. The Student's t-test with p-value= 0.00587 allows us to verify that the home court gives a higher advantage on the game's outcome. Analyzing the home-court advantage at the teams level, we observe that it shows a more substantial influence as the team performs better. Through logistic regression, we verify this behaviour, and we have enough evidence to support that the home-court advantage combined with high-performance teams, can significantly impact the outcome of a game.

Secondly, we examine how the number of rest days affects the outcome of the game. Resting days among 0 and 4 have the most decisive influence on the game outcome as they represent 95% of the observations. Pearson's chi-squared test conclude that we do not have enough evidence to support that the variables are dependent. When we analyze the data on the teams level for away and home wins, a relation was reviled between the 0,1,2,3,4 rest days with the home wins and with general wins but not with the away wins. Therefore we conclude that we have some evidence to indicate that the rest days and home game outcome show some dependance.

Comparing our findings with the literature review, we can conclude that a home team's probability to win has been reduced to 56% for the season of 2014-2015 as it used to be at 65% and fall to 60% in 2011, showing a steady decline behaviour. Moreover, the factors affecting the outcome of the game, like travelling and a home-court advantage, also appeared to be significant for the

2014-2015 session. The relation between the rest day and the game outcome might be affected by the improved travelling conditions allowing player to be more efficient with less rest days, but it still impacts the game outcome.

6 Reference

Courneya, K.S. and Carron, A.V. (1992), "The Home Advantage in Sport Competitions: a Literature Review," Journal of Sport and Exercise Psychology, 14, 13-27.

Harville, D.A., Smith, M.H. and Rubin, D.R. (1994), "The Home-Court Advantage: How Large Is It and Does it Vary from Team to Team?," The American Statistician, 48, 22-29.

Schwartz, B., & Barsky, S. F. (1977). The home advantage. Social Forces, 55(3), 641-661.

Carron, A., Loughhead, T., & Bray, S. (2005). The home advantage in sports competitions: Courneya and Carron's (1992) conceptual framework a decade later. Journal of Sports Sciences, 23(4), 395-407.