MAT022 NBA Report

Rory Tisdall

09 February 2021

Abstract

This report displays a range of descriptive and inferential statistical methods used to analyse and interpret shot data from the NBA season of 2014-15. A particular focus is taken comparing 2-point and 3-point shots, where it is obsverved that the time the ball is in a shooting players hands has an impact on determining the outcome of a 2-point shot. It is found that players are more likely to score a shot if it is taken on their home ground. Using linear regression, a relationship is found, linking the accuracy of a player's 3-point shots and the proportion of 3-point shots out of total shots taken by that player.

Contents

Introduction

The data set contains information on data of shots taken during the 2014-15 season of the National Basketball Association (NBA). The NBA is men's a professional basketball league in North America and is considered to hold the best standards of basketball performance of any league in the world. The 2014-15 season was won by the Golden State Warriors (GSW), with their player, Stephen Curry, awarded the title of season MVP. The data set consists of 128,069 shots taken as observations of 23 variables which describe various attributes of each shot taken. The variables are:

```
## [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"
```

The data set contains shots taken from 281 players, from all 30 teams in the NBA. Each observation of the data set consists of information about the shot; i.e. the distance it was taken from the net, variables which describe the current state of the game at the time when the shot was taken; i.e. the game period, and general information about the game; i.e. The name of the team playing at home. This study takes a focus on the shot type, 2-point and 3-point shots. The shot type is determined by where the shot is taken in relation to the 3-point shot arc, a line that marks 23 feet and 9 inches from the net. 2-point shots are taken within the arc and 3-pointers outside, hence 3-point shots are considerably more difficult to score.

Comparison of 2-point and 3-point shots

This section covers the effect of the duration that the ball is in the shooting player's hands (variable `TOUCH_TIME`) has on the outcome of scoring a shot (variable `FGM`) and the effect that playing on the home ground (variable `LOCATION`) has on the outcome of a player scoring a shot. Both of these investigations are split into 2-point and 3-point shots and any effects found are compared between the two cases. Figure 1 demonstrates that the frequency of 2-point shots is considerably higher than that of 3-point shots. When carrying-out statistical tests, an equal number of random samples are taken from both variables to keep a balanced sample size.

Touch time and shot success

Figure 2 shows the distributions of both shot types from 20,000 randomly selected samples. The mean touch time for 2-point shots is over a second greater than that for 3-point shots, this is likely due to the shooter dribbling closer to the net before the shot is taken.

```
## [1] F-test to compare variances, P-value=1.0000
```

The results of the F-test show that the true ratio of variances between touch time for 2-point and 3-point shots is not less than 1, therefore the variance of 2-point shots is greater.

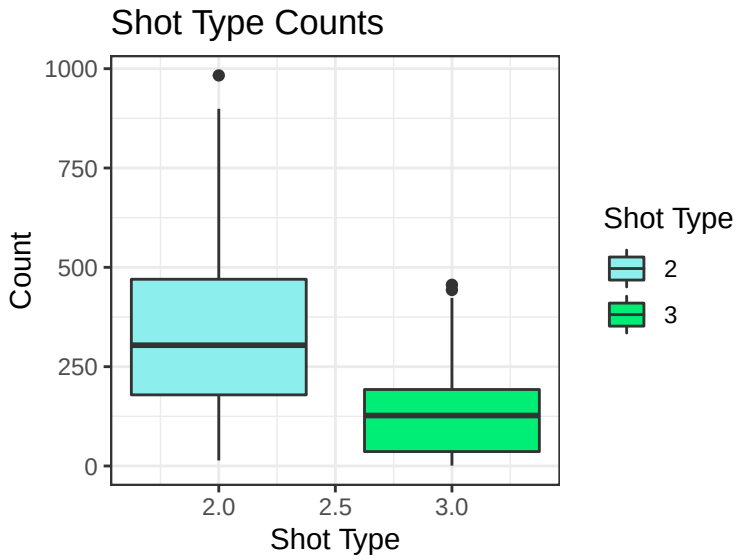

Figure 1: Boxplot to show the number of 2-point and 3-point shots taken by each player.

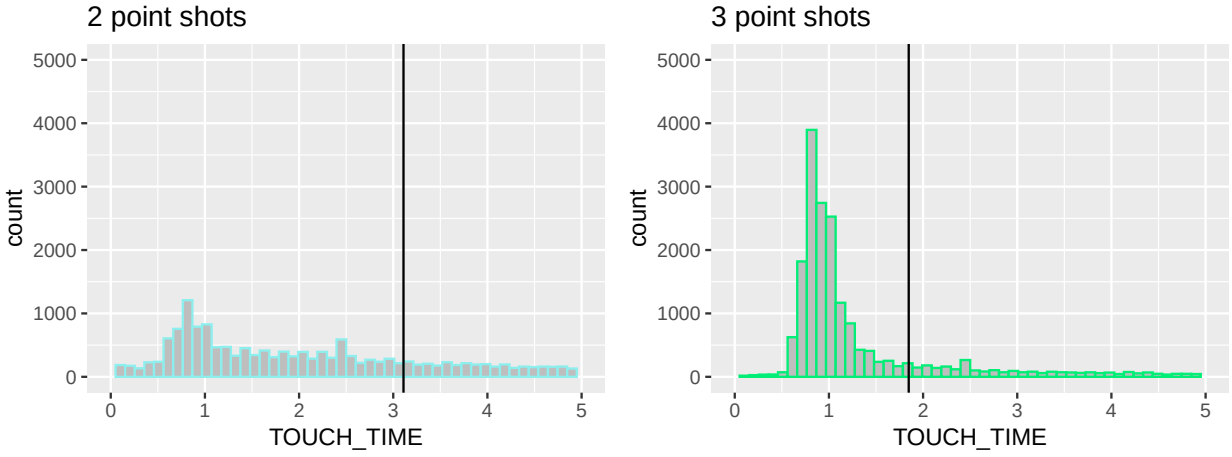

Figure 2: Histograms of touch time for 2-point and 3-point shots including the mean.

```
## [1] Analysis of variances of FGM and TOUCH_TIME

## [1] AOV for 2-pt shots, P-value=0.0000

## [1] AOV for 3-pt shots, P-value=0.0970
```

The results of the one-way AOV test provide very strong statistical evidence that there is a relationship between touch time and shot success when attempting 2-point shots. Unlike the previous case; the null-hypothesis, there is no relationship between touch time and shot success, cannot be rejected at 99% confidence for the 3-point case. However it can be rejected at 90%, suggesting there is statistical evidence that there is a relationship between the two variables. Sample sizes of 10,000 were used for AOV so normal approximation is realistic.

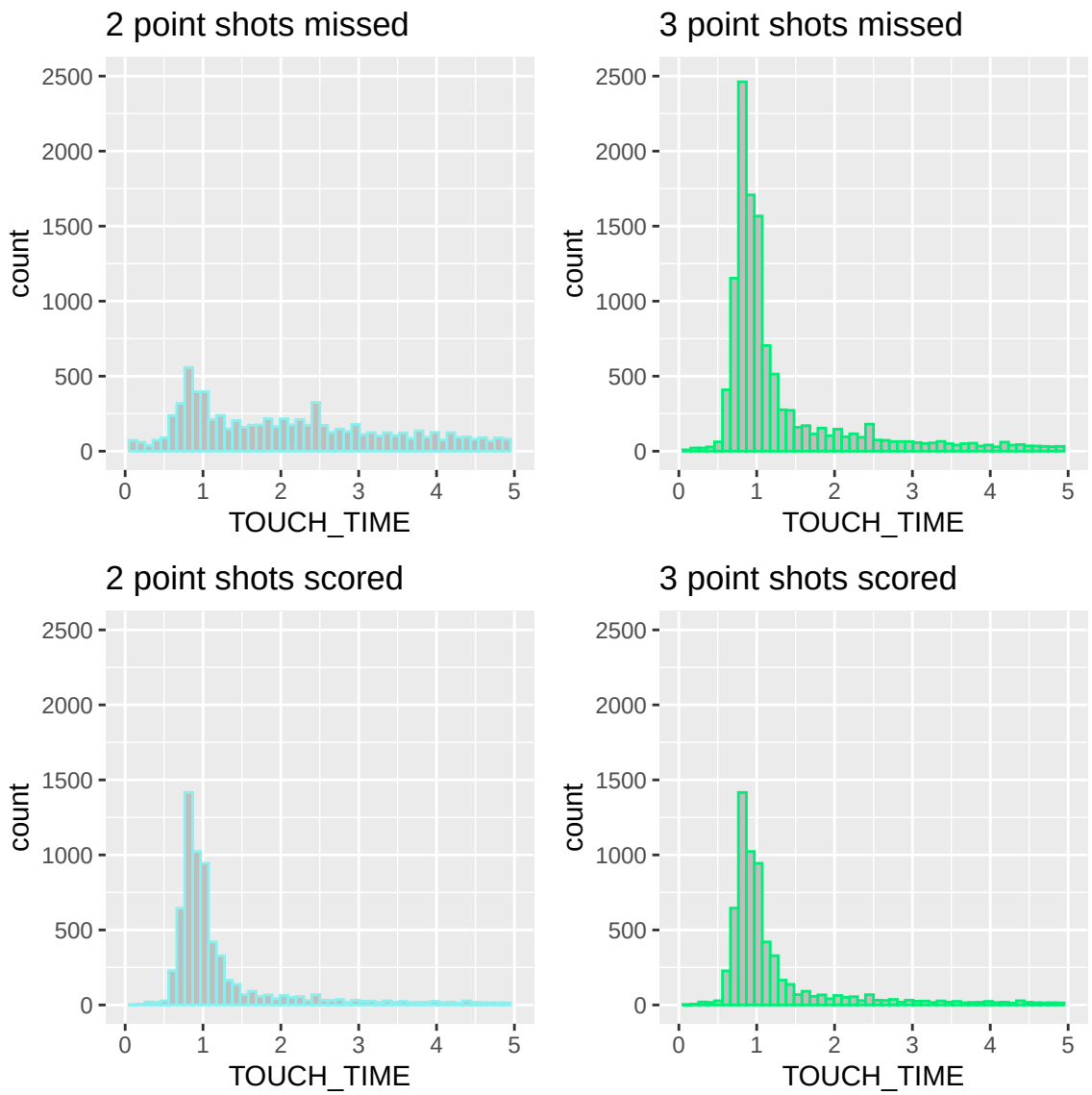

Figure 3: Histograms of touch time with shot outcome as scored or missed for 2-point and 3-point shots.

The relationship between touch time and shot success with 2-point shots can be seen on figure 3, the data suggests that shots taken after 1 second are more likely to miss than those taken before and few are scored after 2.5 seconds, this could be due to defenders having more time to close down the attacker. Few shots are scored before 0.5 seconds, likely because the shot has been rushed. Both observations cannot be seen on the 3-point shot histograms, the shape of the shots missed and shots scored graphs are similar. The reason for this could be that most of the defenders are concentrated within the 3-point arc, marking other potential attackers rather than closing down an attacker who is not considered such a risk.

Location and shot success

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$LOCATION and df$FGM
## X-squared = 8.3565, df = 1, p-value = 0.003843
```

With a p-value of 0.004, the null hypothesis is rejected with 99% confidence and therefore there is very strong statistical evidence of an association between playing on the home ground and the outcome of a player's shot. By splitting the data into two sets, one containing only 2-point shots and the other 3-point shots, a comparison of such an association can be made against the type of shot.

```
## [1] Chi-Square test of LOCATION and FGM
```

```
## [1] Chi-square for 2-pt shots, P-value=0.0205
```

```
## [1] Chi-square for 3-pt shots, P-value=0.0249
```

With p-values below 0.05, the null hypothesis is rejected at 95% confidence for both cases. Although by only a small amount, the p-value for 2-point shots is smaller than the p-value for 3-point shots, suggesting that the relationship of association is stronger when considering only 2-point shots rather than only 3-point shots.

Predicting shot outcome

The shot outcome is binomial and therefore logistic regression is the appropriate method to use to predict this variable. The independent variables: `LOCATION`, `PTS_TYPE` and `TOUCH_TIME` are used to predict the dependent variable `FGM`.

```
## [1] "Model coefficients:"
```

```
## (Intercept)  TOUCH_TIME   LOCATION1    PTS_TYPE
##  0.04856417 -0.03829731  0.03386639 -0.59294448
```

```
## [1] "Model p-values:"
```

```
##   (Intercept)    TOUCH_TIME     LOCATION1      PTS_TYPE
##  2.086331e-02  4.343576e-24  9.802269e-02 1.488770e-174
```

The coefficients suggest: a longer touch time decreases the probability of scoring a shot, the shooting player being at their home ground increases the probability of scoring a shot, and shooting a 3-point shot over a 2-point shot decreases the probability of scoring. The size of the coefficients for `TOUCH_TIME` and `LOCATION1` are very small, indicating the change in probability is also very small. The magnitude of the coefficient for `PTS_TYPE`: -0.6, means that there is a considerable decrease in the probability of scoring when shooting a 3-point shot, as previously seen in this report. The results of all three coefficients are strongly supported by statistical evidence, suggesting the relationships are true and not by chance.

```
##
##                Pred missed Pred scored
##   True missed        19313         3846
##   True scored        12204         4637
```

```
## [1] F1 score for missed shots = 0.7065
```

```
## [1] F1 score for scored shots = 0.3662
```

The model does not perform well at predicting scored shots given by the F1-score: 0.36. The model chooses with greater favour to predict a shot as missed than as scored, hence this model is not very accurate. The main reason for this imbalance is likely due to only 49% of 2-point and 35% of 3-point shots are scored.

Do players with high 3-point shot accuracy tend to take a higher proportion of 3-point shots?

This section takes a focus on each player in the data set, rather than each shot taken as previously. For this, a new data set has been created, having a row for each player with select variables made from aggregating the variables from the original data set. The new variables describe the following for each player: the proportion of 2-point shots taken out of all shots, the proportion of 3-point shots taken out of all shots, 2-point shot accuracy (mean `FGM` for 2-point shots) and 3-point shot accuracy (mean `FGM` for 3-point shots).

Table 1: Mean statistics calculated from all players in the data.

Variable_name	Mean_value
Proportion of 2-point shots	0.7269494
Proportion of 3-point shots	0.2730506
2-point accuracy	0.4839363
3-point accuracy	0.2876193

The values in table 1 suggest that players tend to take considerably more 2-point shots than 3-point shots, and the mean 2-point shot accuracy is considerably greater than the mean 3-point shot accuracy. These stats have led to the investigation, do players with high 3-point shot accuracy tend to take a higher proportion of 3-point shots? This will be investigated by applying least-squares regression to the variables: `proportion_3pointers` and `MEAN_FGM_3_pointers`, with each sample representing a player.

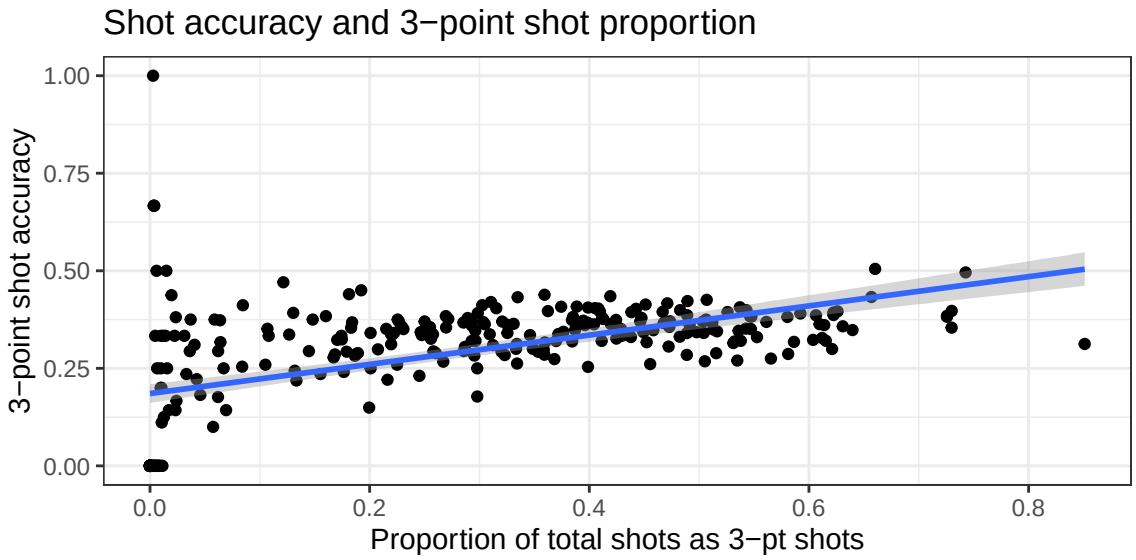

Figure 4: Scatter plot of 3-point shot accuracy against the proportion of total shots as 3-point shots. Notice that there are a number of players who have not taken or scored a 3-point shot and some extreme outliers.

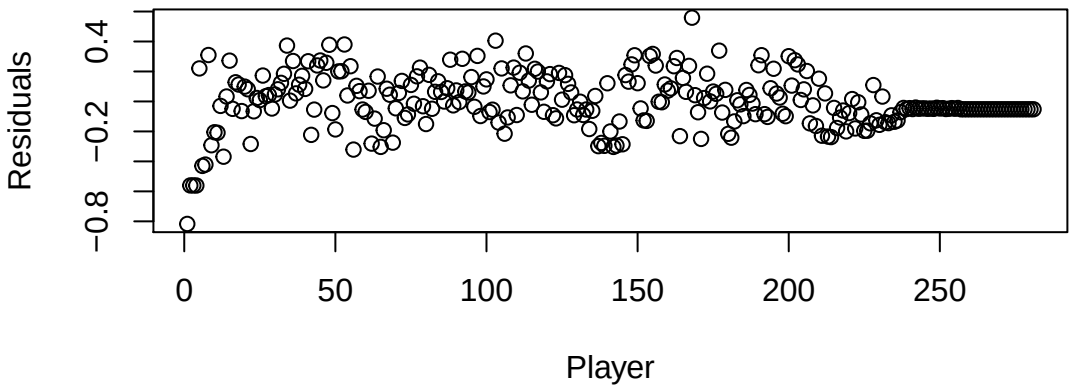

Figure 5: Scatter plot of the residuals from figure 4.

The group of residual points corresponding to roughly the last 50 players all have a residual of 0, while the residual values vary for the rest of the players.This tells us that the regression shows considerable heteroscedasticity, which means that the spread of variances of the residuals are not constant. These players are likely those that have never taken or scored a 3-point shot. One of the assumptions of least squares regression is that the residuals show homoscedasticity, and hence this model is not reliable.

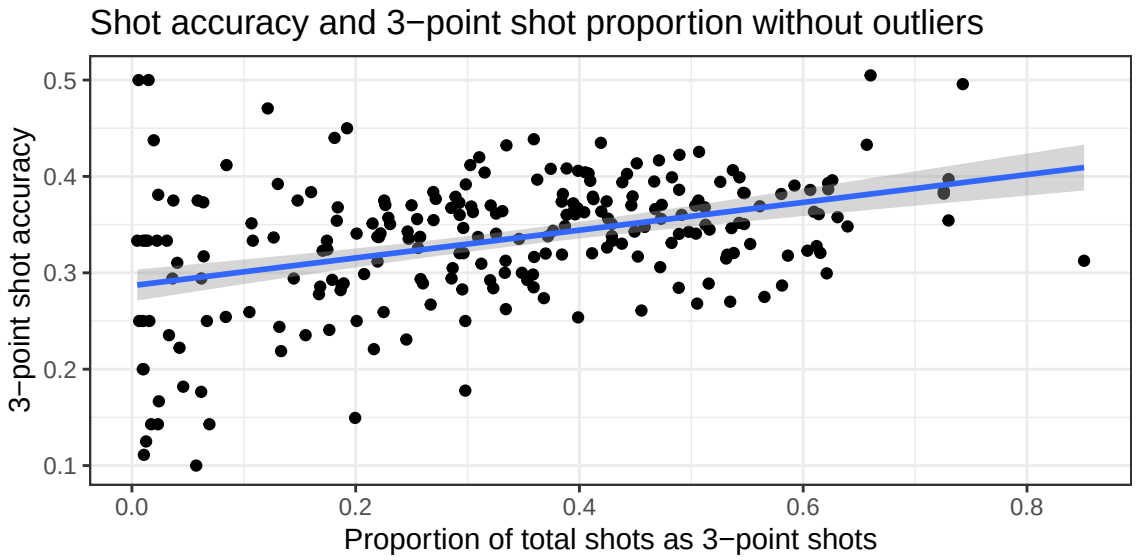

Figure 6: Scatter plot of 3-point shot accuracy against the proportion of total shots as 3-point shots. Outliers and players that have not taken or scored any 3-point shots have been removed.

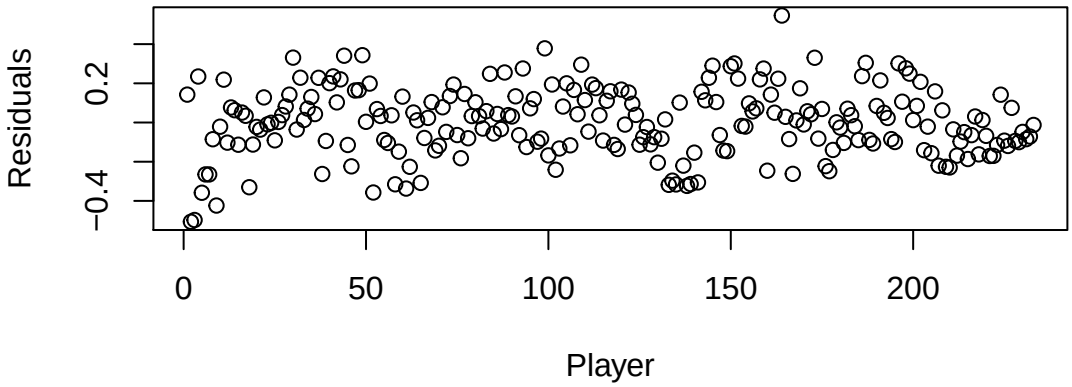

Figure 7: Scatter plot of residuals from figure 6.

The effects of removing outlier players and players which did not score or take any 3-point shots can be seen from figure 7. There is no longer a large group of players with an equal residual value. The model now obeys the assumptions of least-squares regression, allowing reliable interpretation of the results.

```
## [1] "Linear regression model descriptives:"
```

```
##                          Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)         -0.04076046 0.05700152 -0.7150767 4.752837e-01
## Mean_FGM_3_pointers  1.10644607 0.16717305  6.6185673 2.505195e-10
```

```
## [1] Correlation coefficient of X and Y=0.3993
```

The p-value given by the regression is very small which allows the rejection of the null hypothesis, there is no relationship between 3-point shot accuracy and the proportion 3-point shots taken out of total shots. There is moderate positive correlation between the variables, described by the correlation coefficient: 0.4. This model provides strong statistical evidence supporting the statement, players with higher 3-point shot accuracy are more likely to take a higher proportion of 3-point shots out of total shots than players with a lower 3-point shot accuracy. A relationship between these variables is not surprising, generally the more successful somebody is at something, the more they do that something. In the case of basketball, scoring a 3-point shot awards more points than a 2-point shot, hence they are more rewarding for the player and their team. The reason why the correlation between the two variables is not stronger could be because 3-point shots are more difficult to score than 2-point shots, causing players to choose to dribble closer to the net and take a 2-point shot, or pass to another player who is closer to the net.

Conclusion

A range statistical methods have been used in this study to investigate differences between the two shot types, the relationship of other variables with the shot types and the tendency some players have to take 3-point shots over other players. The statistical methods used are: F-test, one-way ANOVA, chi-squared test, logistic regression and linear regression. The main findings of this study are: 2-point shots taken after 1 second of touch time are more likely to miss than those taken between 0.5 seconds and 1 second; this relationship was not observed for 3-point shots, players are more likely to score a shot if they are playing on their home ground, the outcome of a shot was not accurately predicted with logistic regression, players with higher 3-point shot accuracy are more likely to take a higher proportion of 3-point shots out of total shots taken than players with a lower 3-point shot accuracy; this finding is particularly interesting, and it would be beneficial to carry out further research on the matter with a larger sample size, such as data from multiple seasons of the NBA.