Investigating vital factors of NBA

Cardiff University

Foundations of Statistics

MAT022

Individual Coursework

By

Mourouzidou Elisavet

Student Number: C2027059

09 February 2021

Abstract

NBA(National Basketball Association) is the men's most popular and one of the most competitive sports in the world. Insights deriving from analyzing such sports, would be mutually beneficial for players and trainers in terms of upgrade the organization and optimize preparation. As NBA is in the spotlight of interest, lots of researchers are analyzing data regarding this sport and provide thoughtful results. Our study investigated initially that the best players have some tendencies that other players seems to ignore, which helped them enhance their performances. In addition, the Chi-square test of association and a logistic regression model revealed us some factors which affected the shot result. Finally, we implemented some tests regarding essential correlations in the NBA data set which resulted to highlight factors that are connected during a game.

Contents

1 Introduction

Every game in the National Basketball Association is very demanding. Every autumn a new season initiates for the NBA until the summer when the playoffs end. While players are constantly training, the NBA fans all over the world wait eagerly to count points and wins in favor of their favorite teams or players. Players need to take decisions instantly and a game can be determined for that one decision. There is a variety of important information in a NBA game and infinite data. Analyzing a NBA data set can be very challenging in terms of statistics as there are many variables to consider. However every result can reveal new techniques for training and preparing the players, but also could be helpful for players to understand the different play styles from other athletes.

Initiating from the descriptive analysis, where we are about to explore and understand our data, we took our first inspirations that we will investigate in inferential analysis, which consists of three main parts. As far as inferential statistics is concerned, we will investigate some of our speculations across the population and not just the given data set. Interpreting results from a data set to the whole population is very difficult but interesting at the same time. In our research we will make some assumptions based on the seven attributes: the touch time that players need before shooting, the distance from the basket and from the closest defender, the dribbles, the period, the type of shots and the shot result.

2 Background

NBA demands continuous effort and discipline from the players in every game. The main aim of every player is to keep his performance as high as possible and manage to score before the end of a match in order to get the victory for his team. Each time a player scores a shot, his team's total points and his personal score increases by one, two or even three points depending on the distance before shooting. Every individual who likes NBA has its favorite athlete. It is evident that mostly the best players in terms of skills are the ones who gain the most of the spotlight. But, how do we define the best player?

Firstly, we have created an extra column named "Success_rate" illustrating a percentage of success for each player. We used it as a measure to distinguish players with best and bad performances. For our report we created a new column with unique number 1 and we added it, considering the players in order to find out the total number of shots. Then we divided the total points that a player achieved with his number of shots. As a result, that measure "Success_rate" takes into consideration not only the points that players have collected but also the number of times he needed to achieve these points in order to calculate each player's efficiency.

We initiated our research with the descriptive analysis, where we aim to explore and visualize the provided data set. Some histograms will illustrate the distributions that our attributes follow, while box-plots demonstrate not only the extreme observations, but also other significant information about the features.

Competitive as NBA is, and irrespective of the struggles that players are dealing with in order to achieve, it is worth mentioning that some players have incredible performances while others seems to have difficulty in performing and scoring. What are these players doing differently? Can we assume that they shoot more often or just that they shot closer to the basket? A report Shaoliang Zhang, Miguel-Angel Gómez, Hongyou Liu, Bruno Gonçalves & Jaime Sampaio, 2017[1] which identified technical and physical performances of basketball players according to playing position in strong and weak teams, inspired our research, as their results showed that technical performances differed between players of strong and weak teams. The strong teams had different techniques as they covered shorter distances and lower speeds than players from the weak teams. In our case, we are about to investigate our assumptions that the best players have different techniques from the weaker players. For that purpose, we extracted two samples from our given data. A

sample regarding the most skillful players and a second considering some others with the lower performances. We contacted some F-test (tests of the equality of two variances) and Welch's t-tests trying to reveal possible combinations that best players may do differently.

Previous findings Elia Morgulev, Ofer H.Azar, Michael Bar-Eli, 2019[2] exhibit that there is a psychological momentum in the specific setting of overtime in basketball games and that maybe there is an impact on their performances. Drawing upon these findings, we decided to test if in our data set the players performances are affected by some psychological reasons such as anxiety when the game approaching to its end. As a result, it is reasonable to assume that there is a possibility to make different decisions in terms of the type of shot and this might cause a negative or positive impact on the shot result. In our report we used the Chi-square test of association to locate if there is any connection between type of shots and shot results with the different periods in a game.

The more efficient a player is the more contributes to his team victory, which is the main concern in an NBA game. Hence, the result of each players shots is one of the major concerns in NBA. What does influence more the shot result? Continuing with the inferential statistics, we introduce our numerical variables as predictors into a logistic regression model, in order to describe the data and to explain the relationship between the attribute shot result and the explanatory variables. A relevant literature Victor Okazaki, Andre Rodacki, Miriam N.Satern, 2015[3], gave us that idea, as in that report they investigate some variables that influence shooting.

At the last part of inferential analysis but as important as the previous parts, we will test some quite large correlations between our variables with Pearson and Spearman tests. Afterwards, we will separate the two most correlated variables based on the shot result (made/missed) and the we will test their correlation again.

3 Descriptive Analysis

3.1 Sample with best performances

For the purpose of our analysis, we extracted data ($n_1 = 396$) regarding the five best players based on the highest Success rates. It is remarkable that scoring percentages were 51% and 64% regarding the five players with the highest performances and the best of them respectively. Moreover, the average distance from the basket when taking a shot was 7.76 feet (SD=8.32) which implies attempts that counts for two points. Likewise, the mean in terms of touching time that players need before shooting was approximately 1.31 seconds (SD = 1.12).

Measures of central tendency and spread of our data can be found in Table 1.

Table1	Touch Time(s)	Closest def Dist(ft)	Shot Distance(ft)	Success Rate
Mean	1.31	3.36	7.76	0.51
Median	0.9	2.6	3.8	0.42
Variance	1.27	7.39	69.21	0.02
SD	1.13	2.72	8.32	0.16

Inspection from the histograms considering continuous variables, showed that touch time (left histogram) is distributed exponentially, while the distance from the closest defender (the histogram in the center) seems like a non-symmetrical normal distribution with right-skewed curve. Hence, we can conclude that these data may follow the lognormal distribution. The right histogram is regarding shot distance, depicting an irregular shape and does not seem to follow a known distribution. In addition, some general observations are that the mean is not at the center of the histogram, while the median is close to zero, meaning that most of our observations are between zero and ten.

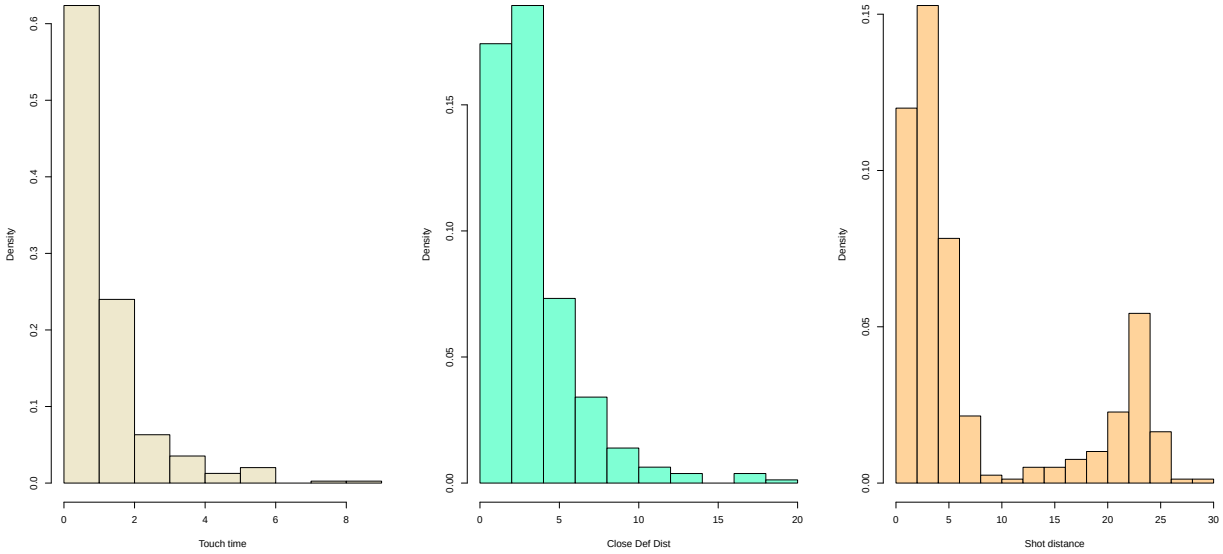

Figure 1: Histograms of `Touch time` (left), `Distance from the closest diffender` (center) and `Shot distance` (right).

Interesting information stem from the one discrete variable, which in our data set is the dribbles that players did before shooting. The period in a match, the points that the players attempted to achieve and the results of these attempts are the categorical variables. Below there are 4 box plots considering the previous mentioned variables with some of the continuous features. The illustrated extreme observations in these figures are not considered as outliers but as different patterns during a game.

Moreover, we observed some significant facts comparing the sample of best players with the first given data set, such as that the best players do not play overtimes since in the new extracted sample we have only four periods. Furthermore, it was a considerable decrease in terms of dribbles as were reduced in a maximum of seven from thirty-two.

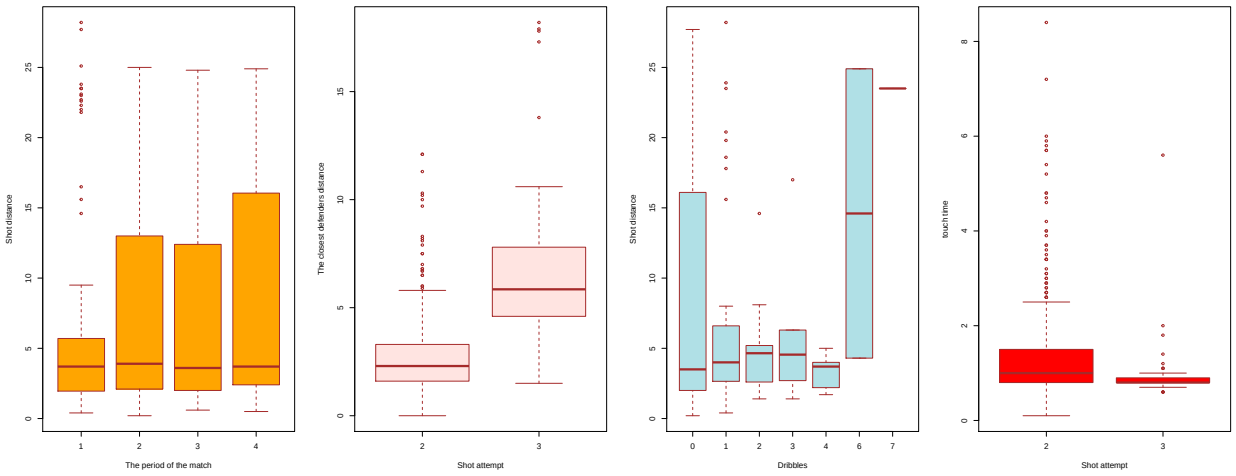

Figure 2: Boxplots regarding `Shot distance` as function of `Period` (first), `Distance of the closest deffender` as function of `shot attempts` (second), `Shot distance` as function of `dribbles` (third) and `Touch time` as function of `Shot attempts` (fourth).

In the first box plot, we noticed that during a match, there is a rise regarding shot distance and the fourth box plot depicts a possible connection in touch time and the points type, as we can see that there is a reduction in touch time when it comes to three points shot.

3.2 Sample with low performances.

The second sample that we extracted from our initial data set, contains the three players with the lowest performances $(n_2 = 2.654)$. Regarding this sample we mainly focused on two attributes, and we found some of the measures of central tendency and spread in order to understand the data (mean ± SD). The first attribute is the Touch time (4.11 ± 3.52) and the second is the Shot distance (14.17 ± 7.74). Another interesting feature in players with low performances is that the mean in success rate was 0.08 (0.43 less than success rate in best players). We wanted to highlight that the player with the worst performance had lost 62% while all three weakest players had lost 58% percentage of the attempted field goals, as illustrated in the pie charts in appendices section (Figure 6).

Table 2	Touch Time(s)	Shot Distance(ft)
Mean	4.11	14.17
Median	3.4	15.4
Variance	12.38	59.91
SD	3.52	7.74

4 Inferential analysis

4.1 Normality

In the first part of inferential analysis we are about to investigate the normality in our data. Firstly, our sample regarding the best players consists of n_1=396 observations whereas the size of the second sample is n_2=2654. By observing our data and the histograms above for every variable, attributes do not follow the Gaussian distribution, but in such sufficient large samples with finite mean and variance (as we mentioned in descriptive statistics), we can assume that the Gaussian approximation is realistic, based on the Central Limit Theorem which conditions for that speculation were satisfied.

However, we implemented another test in order to confirm our assumption. Before initiating with normality tests in our attributes, we set $\alpha = 0.05$ which means that every time a statistical test resulted to a p-value equal or lower than our significance level, and after consulting the critical region, the null hypothesis (H_0) can be rejected in favor of (H_1) hypothesis, with a maximum of 5% probability of a false positive (Type I error). The Shapiro test (p= $2.2 * e^{-16}$) was used for that purpose and as the central limit theorem refers, indeed there are strong statistical evidences that our data were normally distributed. The confirmation of the normality assumption is the key feature to continue with our investigation, since we are about to search for other speculations, using tests which demand normality.

4.2 Hypothesis tests for differences between best and weak players

We are about to investigate two variables: the touch time before shooting and the Shot distance in our two samples (Players with high and players with low performance). Our purpose is to find if the players with higher success rates keep the ball less than the other players and if they also prefer to shot when they are quite close to the basket. Were these some of their techniques which increased their success or just random observations?

4.2.1 Variable touch time

Starting with the variable **touch time**, firstly we intended to analyze the variances using the F test. We set the null hypothesis that the **variances** between the two samples were equal, while the alternative implies that there were different. We observe that 1 does not belong to the 95% confidence region for the equality of variance test. The obtained p-value($< 2.2 * e-16$) was very low and thus we can conclude that there are statistical evidences suggesting that the variances between the two samples could not be equal.

Table 1 and table 2 shows that the **mean** of the time that players need before shooting is approximately 1.3s and 4.1s for the players with higher and lower performances respectively. We observed a difference between means and we are about to search if we can assume that best players intend taking a field goal rather than keeping the ball without passing it. In order to investigate the previous mentioned assumption, we implemented the Welch two sample t-test for the means setting as null hypothesis equality between the means and as an alternative we set that best players have lower mean. Results indicated that we have strong statistical evidences to reject the null hypothesis($p < 2.2 * e-16$) in favor of the alternative and thus we can suggest that players with high performances seems to avoid consuming time by touching the ball before shooting. Maybe this is a possible pattern which allows them to achieve a higher performance in terms of scoring.

4.2.2 Variable shot distance

Alike, for the **shot distance** we tested the variances and the means in the two samples. Starting with the F-test for the **variances**, we observed that 1 belongs to the confidence region at the significance level of $\alpha = 0.05$ and thus we can conclude that there are statistical evidences relatively high p=0.052 suggesting to retain the null hypothesis that the variances of shot distances could be equal.

The **mean** regarding shot distance was 7.75ft and 14.17ft for players with the best and low performances respectively. We used a t-test by setting as null hypothesis $H_0 : \mu_1 = \mu_2$, with the alternative $H_1 : \mu_1 < \mu_2$. The results revealed that there are significant evidences (p=$2.2 * e^{-16}$) suggesting to keep the alternative hypothesis instead of null. Hence, we can conclude that $\mu_1 < \mu_2$ which interpreted as that maybe the best players have their technique and do not prefer taking the risk of shooting for more points and from larger distances.

4.3 Tests for association

In the previous part of our study, we revealed some possible techniques that the best players may use, which may contribute to increase their performances and make them more efficient. In the next part of our analysis, we will focus only on the best players and we will test for some other factors that may be associated with their shot result.

4.3.1 Chi-Square test of association

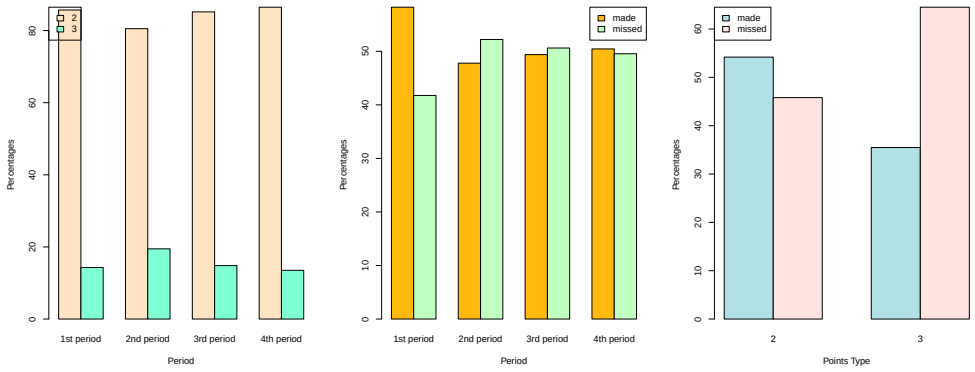

Figure 3: Bar charts regarding `Points type in every period`, `Shot result in every period` and `Shot result considering points type`

We are about to test three assumptions. Before reaching to the shot result we will search if the period is associated with some factors in an NBA game. Our first test is regarding the best players decisions in terms of shooting with the period, while the second test provides information about the shot result with the periods during a game.

We argue that as close as we get to the end of the game, best players may feel anxious and attempt more shots of two points rather than three, and that as the time elapses may affect the result of their attempts. The third assumption, is about best players abilities. To further explain, we are about to investigate if there is any relationship between the shot type that the players intend to shot with the shot result. Finally, rushed by our three tests of association we can conclude if there are any significant two way interactions between the categorical variables in our data set.

For those reasons, we implemented three different X^2 Chi-squared test of association to reveal if these variables are somehow connected with each other. The null hypothesis supports that there is no relationship between two categorical variables. We observed for the first assumption that the Pearson's Chi-Squared Statistic, $X^2 = 1.8028$, corresponding to a p-value = 0.6143, and in our second assumption $X^2 = 2.4642$, corresponding to a p-value = 0.4818. Thus there are no statistical evidences implying that shot type and shot result are influenced by the period in a match. Our speculations about playing in a safer way and feeling anxious does not seem to be valid. To be more specific, athletes may play at the same way the whole time in a game and not be affected by the period.

However, our third assumption which considered type of shooting with the final result, the X^2(6.5956) test of independence with p-value = 0.01022 implies significant evidences to reject the null hypothesis in favor of the alternative, and thus there are evidences suggesting association between points type and the shot result. Inspired from that result and based on the (figure 3), we are allowed to hypothesize that scoring three points is more challenging but at the same time vital than scoring two points and there is a higher risk when even best players aim for the three points.

4.3.2 Logistic Regression

Previously we were conducting tests in order to find associations between sets of categorical factors which may influenced from the period and may be connected with the shot result. Now we are about to investigate how some numerical factors such as touching time of the ball, dribbling and shot distance from the basket influence the predicted variable which is the shot result. Therefore, we are about to create a logistic regression model with all four continuous variables and see how relevant are with the input, and how they contribute to our model. Before continuing, we wanted to highlight that we consider that the outliers were not measurement errors but could potentially represent extremely high or low attempts.

Afterwards, before setting our model we used Variance Inflation Factor (VIF) test for detecting highly correlated predictor variables. VIF results fluctuated between 1.7 and 2.07 indicating lack of multicollinearity among our predictors.

Table 3	shot_dist	close_def_dist	touch_time	dribbles
VIF values	1.71	1.72	2.08	2.06

In the model we used as predictors: the shot distance, the closest defender distance, the touch time (three continuous attributes) and one discrete variable the dribbles to predict the shot result (made or missed). Below we present some more information regarding our specific type of Generalized Linear Model, a Logistic Regression model.

Before reaching a conclusion we are about to explain how our model works and its capabilities. Deviance residuals represent the square root of the contribution that each data point has to the overall Residual Deviance. Even if Deviance Residuals median is 0.8006 and they are not actually centered on zero, they are roughly symmetrical, as the minimum value is -1.5176 and the maximum value is 2.0964. That high numbers indicate that these residuals may be caused by outliers.

Afterwards, the coefficients column correspond to the following model: $Shot\_result = 0.69446 - 0.08106 * shot\_dist + 0.08929 * close\_def\_dist - 0.26804 * touch\_time + 0.07240 * dribbles$. To further explain, minus in the "`Estimate`" column means that for every one unit of shot_distance gained, the log(odds) of shot_result = "made" will be decreased by 0.08106, while touch time is interpreted in the same way. On the other hand, regarding the other two variables "close_def_dist" and "dribbles" when they increase, the

log(odds) of shot_result = “made” rises also.

```
##
## Call:
## glm(formula = IsMADE ~ shot_dist + close_def_dist + touch_time +
##     dribbles, family = binomial(link = "logit"), data = data_MADExMISSED)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5176  -1.2075   0.8006   1.0321   2.0964
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.69446    0.23151   3.000   0.0027 **
## shot_dist      -0.08106    0.01721  -4.709 2.48e-06 ***
## close_def_dist  0.08929    0.05048   1.769   0.0769 .
## touch_time     -0.26804    0.14262  -1.879   0.0602 .
## dribbles        0.07240    0.17948   0.403   0.6867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 548.72  on 395  degrees of freedom
## Residual deviance: 515.33  on 391  degrees of freedom
## AIC: 525.33
##
## Number of Fisher Scoring iterations: 4
```

In terms of the crucial attributes, shot_dist variable is very important for our model as it contributes to a significance level of 0.001 with a very small p value ($= 2.48 * e^{-6}$). The next two variables close_def_dist and touch_time do not have a major role in our model but they may influence at a significance level of 0.1. Finally, the intercept of our explanatory variables has quite small p-value=0.0027 indicating that in a significance level of 0.01 the log odds of shot_result=“made” will increase by 0.69446.

To sum up, as the model revealed it is logical to believe that the shots that are made from a higher distance to the basket reduce the probability of succeed. In addition, touch time might have a negative effect regarding successful field goals. At this point, it is worth mentioning that the test in 4.2.1 revealed approximately the same result (since best players might have quite low touch time). Whereas, the closest a defender is to a player who intends to shot the more difficult is for the shooter to score.

4.4 Correlation tests

Some interesting correlation between the factors were revealed such that the touch time was highly and positively correlated (r= 0.708) with the dribbles, alike close defender distance is positively correlated (r=0.614) with shot distance. We are about to investigate if these correlations are statistically important. As we have mentioned the Gaussian approximation is realistic due to efficient large sample size. So, we can say that the results of the Pearson tests will be reliable and trustful.

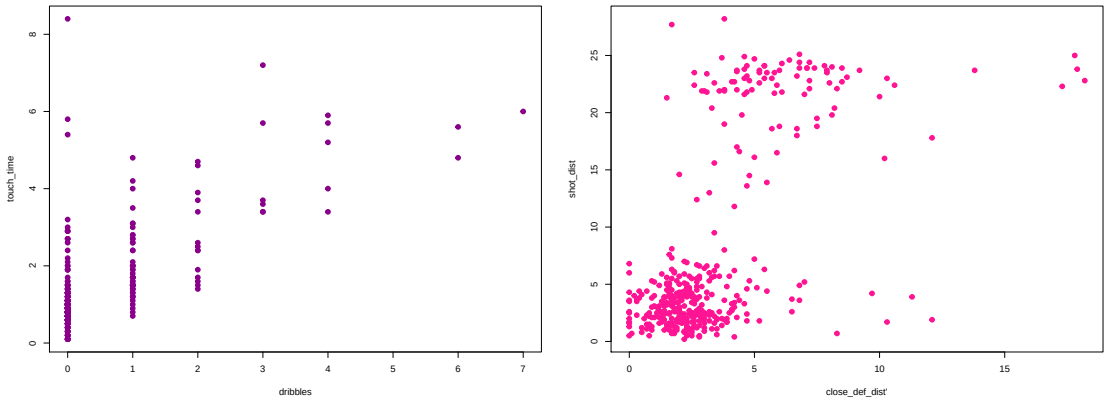

Figure 4: Scatter plots regarding `Touch time` with `Dribbles` and `Shot distance` with `Closest defender distance`

We used two Pearson's correlation tests considering the aforementioned sets of attribute's. Both tests resulted to a considerable small p-value. However, influenced from the outliers in our data set, we decided to use also Spearman tests for correlation and see if there is a difference in their results. Spearman tests indicated also very small p-values. Hence we can conclude that each of these pair of attributes are correlated in a significant level. As expected, it is rational to believe that most of the time that players have the ball they tend to dribble more rather than keeping the ball in the same spot. This fact, contributes to the high positive correlation. Moreover, it is logical that players have the tendency to be as far as they can from the defenders while they are away from the basket when attempting to shot. This practically means that the defenders are close to the basket and at the same time the attackers are approximately on the three point line. In general, defenders constantly approach attackers, while attackers decisions are always according to the defenders position.

4.4.1 Made/Missed Correlation test

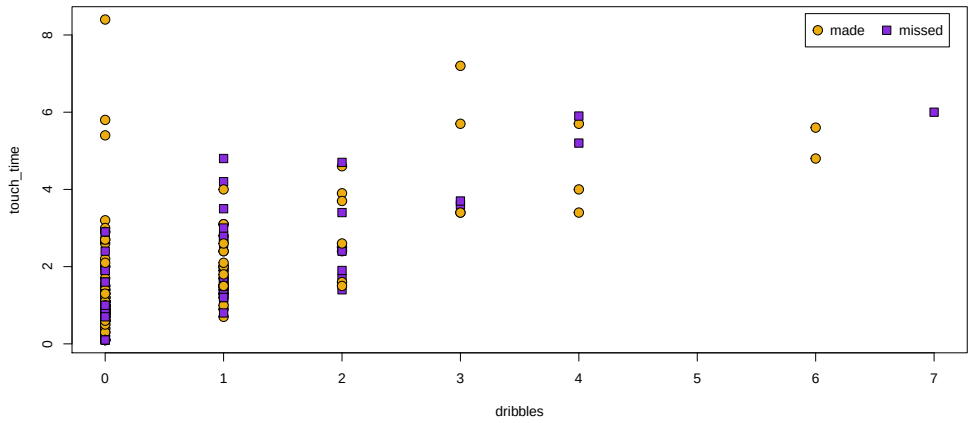

Figure 5: Scatter plot regarding touch time and dribbles in terms of the shot result

Irrespected of the previous correlations it would be more essential to implement some more detailed correlation analysis in terms of the shot result. For that reason we separated the touch time and the dribbles observations

based on when the attempt was successful or not. We used the non parametric Spearman test twice to determine if the assumption of correlation is realistic. On the topic of the test, we can reject the null hypothesis against the alternative in both tests, cause of the very low p-values. Hence, we can safely assume that whether the shot was made or not the linear dependence is reasonable. As a result, to some extent when dribbles increase, touching time rise also and vice versa despite the final result. From that tests we could not distinguish our cases in made or missed, as in both cases the results seem not to change.

5 Conclusion

Overall, the results of each test have rational meaning while were intuitive with someone familiar or not to the terms of basketball and to the concepts of the variables.

Our first remark regarding the differences between players with high and low performances indicated that the difference between means in terms of touch time and shot distance were significant. As a result, our tests revealed some patterns regarding the variety of styles during a game and some decisions that best players make differently. Some further research can be done in comparing attributes between best and weakest players regarding other variables such as final margin and the time that players spent on the court.

It is commonly believe that when time runs out impacts negatively on decisions and on human's efficiency. However, chi-squared tests of association did not reveal a significant connection. Maybe in a different data set or different sport, the period might influence the players contribution. In addition, the third chi-squared test indicated connection between the type and the result of shots. Definitely, that result is logical since as far as a player is from the basket (three points shot) the harder is to achieve a successful shot. We made that assumption stemmed from the chi-squared results but our logistic Regression model remarked the shot distance as the most essential factor for predicting the result of the shot. When shot distance and touch time increase, it is more difficult for the players to achieve a successful shot. Even if the distance from the closest defender and dribbles were not very important predictors, they had a positive impact on increasing the probability of the successful field goal. It would be really interesting to expand our analysis by adding more and different variables to predict the same output.

Last but not least, we implemented tests considering some quite high correlations in our data set. The results showed that these correlations indeed were significant. Afterwards, we continued to a final research separating the two outcomes of the shot result and connected touch time and dribbles. In both cases (made/missed) there was correlation between the variables and we did not reveal any unique connection between them but could be possible be an assumption for a future search regarding some other variables. Furthermore, undoubtedly someone could make an analysis about a linear dependence of two really vital categories in NBA named turnovers and rebounds.

In general, we tested some very specific attributes for the best players but there is a plethora of variables that can influence and might contribute to a player's performance and to the final shot result.

6 References

[1] https://www.tandfonline.com/doi/full/10.1080/24748668.2017.1352432

[2] https://www.sciencedirect.com/science/article/pii/S0167487017307122

[3] https://www.researchgate.net/publication/279180866_A_review_on_basketball_jump_shot

[4] https://stats.idre.ucla.edu/r/dae/logit-regression/

7 Appendices

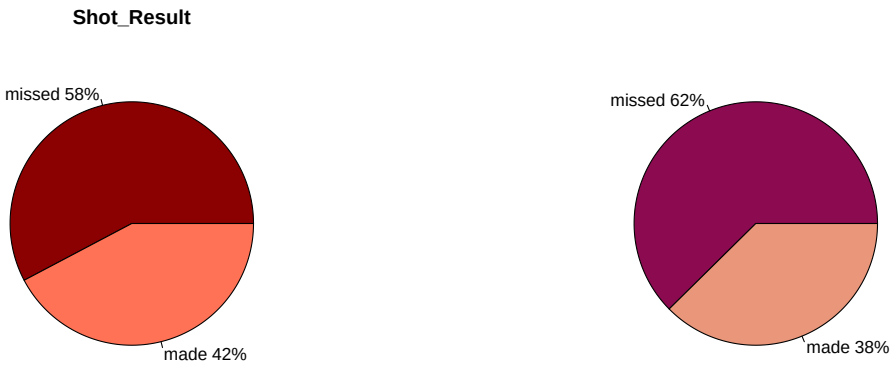

Figure 6: Distributions for the players with low success rates(left). The player with the lowest success rate (right).