A Study of NBA Dataset

Mert Kaya

23 February 2021

Abstract

Our main purphose is to carry out some descriptive and inferential analysis over the nba data set, and to discover some considerable statistical evidences to evaluate the performance of teams and players, and the shoot results. First we start with investigating the performance of teams by looking into the margin scores of each team, and the significant effect of game locations over the win rates of the teams. We display some studies to emphasize the importance of 4 main periods on shoot attempts, successful 3 pointers, scores made in each period and shoot distances. Then we explore the significance of shot clock on successful 3 point shoots. Furthermore, we create a correlation matrix to see the relationship of some features and take advantage of the correlation statistics to perform a multi-linear regression in order to predict the average scores made by the nba players.

Contents

1 Descriptive Analysis of the NBA dataset **2**

2 Nba Team Performances by Game Locations **5**

2.1 Margin Comparison of Teams 5

2.2 The Effects of Game Locations 7

3 Differences Between Periods **8**

3.1 2 and 3 Pointer Attempts Comparison Between 4 Main Periods 8

3.2 Connection of Successful 2 and 3 Pointers with Periods 9

3.3 Location vs Periods for 3 Points Shoot Attempts 10

3.4 Shoot Distances By Each Period 10

4 The Effect of Shot Clock on 3 Pointers **11**

5 Average Score Predictions of Nba Players **12**

6 Conclusion **14**

1 Descriptive Analysis of the NBA dataset

The nba data set is basically consist of every single two and three pointer attempts made by the players and their corresponding detailed features during the matches played for approximately 2 months. Since nba is an intense league, although data was collected for a short period of time,the nba data includes 128069 observations with 23 variables.

The variables given to analyze the shoots are listed below:

```
##  [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"
```

Since the data set has many features, a lot of analysis and inferences can be made, and we are interested to reveal some significant statistics.

The only feature that has missing data is SHOT_CLOCK with 5567 observations and because of that situation, we can omit the missing data only when we are analyzing the shot clock.

```
## [1] 5567
```

Although the data set do not include all the nba players, it has 281 of nba players which is sufficient and it also has 30 all of the nba teams that were playing in the current term. Moreover, the data contain 904 matches.

The nba data is complicated at first sight and it has to be regulated somehow.Thus, in our study, we organize the data to obtain the nba player and team information to use them to get some inferential analysis (see appendix for further descriptive analysis of the data).

Overall, when we analyze the nba teams, we emphasize the margin difference and the importance of match locations over win rates of the teams. We investigate the periods with respect to amount of shoot attempts and investigated the effects of periods on successful 3 pointers. We also examine into the total scores and shoot distances by each period. In a separate title, we observed the relationship of shoot clock with the 3 pointers. When it comes to the player statistics, we create a correlation heat map to observe the connections of some features and finally we perform a regression with some major features to predict the average scores of nba players.

In this section, we will also emphasize some key points of the data to understand the features deeply.

Figure 1, represents the margin statistical information for the NBA teams played in home location. If we examine the first noticeable teams, when we look at CLE home margins scores, the median is much bigger than the mean thus we may state there are some outliers which affected the statistics deeply and these games affected the average score significantly. The variance is the highest among all teams and since, it might support our claim about outliers. When it comes to GSW, it has the hightest margin average and median among all teams for the matches played in home. Since GSW has low variance compared to other teams,

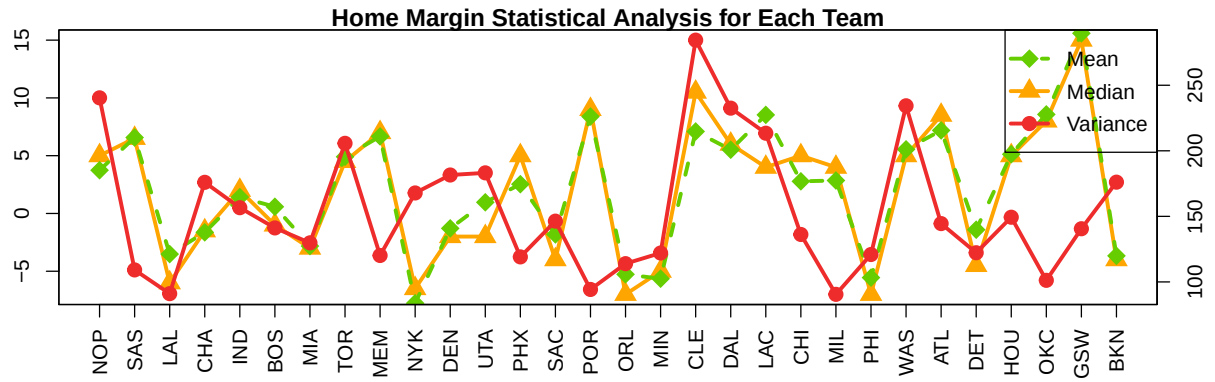

Figure 1: Margin means,medians and variances in Home for the nba teams

their margin values in each home match are more consistent when compared with many teams. But since we have approximately 30 match data played in home for all NBA teams, these comments and statistics might change a lot.

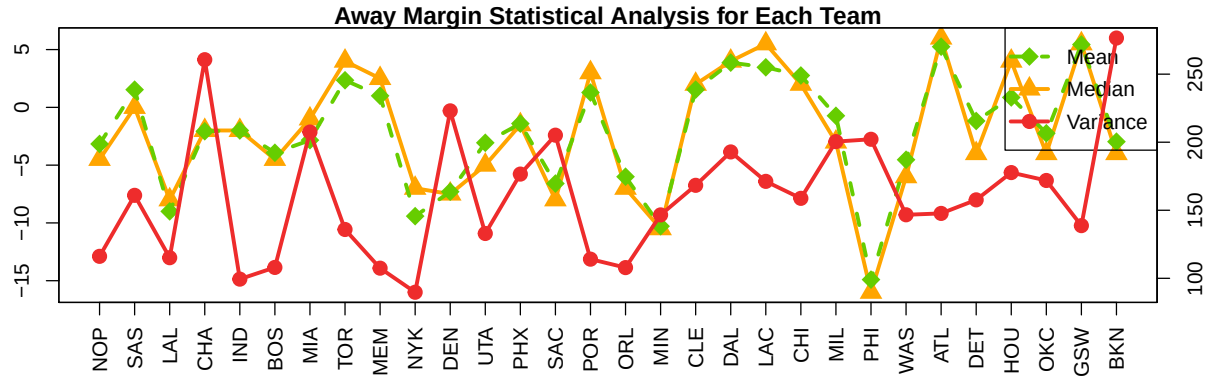

Figure 2: Margin means,medians and variances in Away for the nba teams

In the Figure above, BKN has the highest variance in away matches according to the data, thus we might say their performance is very uncertain in away stadiums. They also have approximately -4 average margin and sligtly lower median than average.Furthermore, PHI has the lowest margin mean and median with reasonable variance. Their performance in away is reasonably worse than the home.

When we look at ATL, they have the highest mean and median with a reasonably low variance. According to these statistics, they are really successful in away matches compared to other teams and their performance is consistent.

The correlation heatmap shown in Figure 3 includes some significant connections between NBA player features. Some reasonable points are:

- Score average is positively correlated with touch time average
- Score average is also positively correlated with dribbles average
- Touch time average and dribbles average are strongly positive correlated
- 2 point success percentage is negatively correlated with shoot distance of 2 points

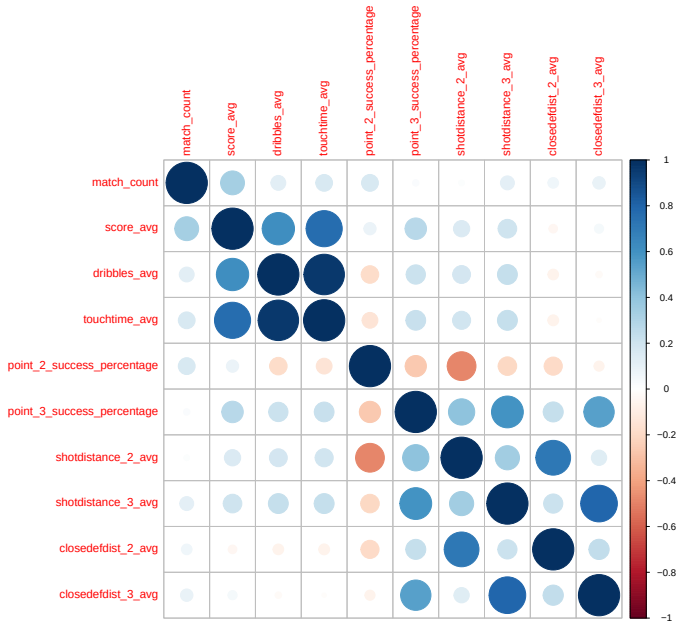

Figure 3: Heatmap for the partial correlation of the average player features in nba dataset

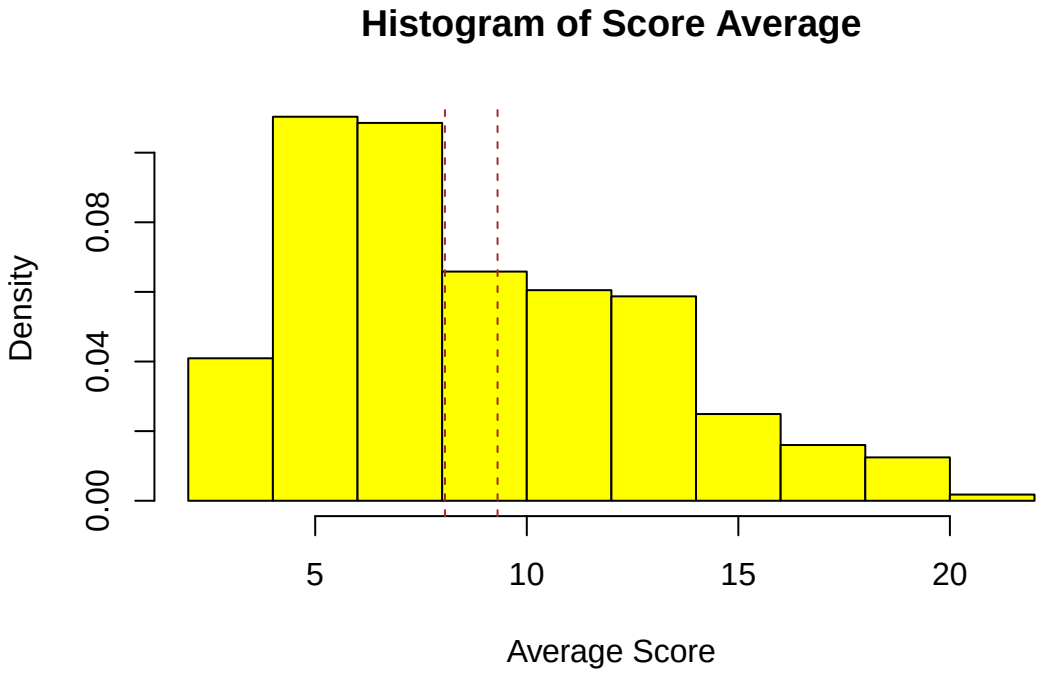

Figure 4: Nba players average score histogram with mean confidence interval of 0.99

In the Figure above(Figure 4), the score average of each nba player can be seen(free throws are not included). As we can see, big majority of players have approximately 5 and 6 points average. On the other hand, when we examine the average confidence interval, we see an interval between approximately 8 and 9.3 points. Perhaps some players with high scores pull up the average significantly.

2 Nba Team Performances by Game Locations

As we watch nba matches, we usually see the high support of the audience for the players during the games and in public, there is a general view that the audience affects the performance of players massively. Thus, in this section, we will try to find some evidences about the effect of game location over the nba teams to support this claim.

2.1 Margin Comparison of Teams

Table 1: Variance tests of the margins of each nba team for locations HOME and AWAY respectively

Test	p.value	Test	p.value
Bartlett	0.6747380	Bartlett	0.4696472
Fligner-Killeen	0.0837703	Fligner-Killeen	0.2325164
Levene	0.6309699	Levene	0.4513080

In Table 1, these 3 test results for the locations of home and away describe no significance to claim that variance of margins are different for each team for the significance level of $\alpha = 0.05$. But since we are applying these tests for all teams, we might ignore the difference of variance between some specific teams.

To investigate the mean margin's of each nba team in the Home and Away locations, let's apply 2 one-way anova tests. Thus we found very small p-values for both of them, then we have strong evidences to conclude that there are differences in margin means for the at least 2 nba teams in both locations.Since we have approximately same amount of matches for each team for both locations, the data that we investigate is stable.However, there are approximately 30 matches in home and away for each team, and such low amount of data might cause wrong results for our anova tests. Thus we should investigate some specific teams.

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## HOME_TEAM    29  25218   869.6   5.577 <2e-16 ***
## Residuals   874 136284   155.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## AWAY_TEAM    29  20765     716   4.447 2.16e-13 ***
## Residuals   874 140738     161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5 suggests that the mean interval of home margin of GSW might be higher than other teams. When it comes to matches in away, PHI may have the lowest margin mean in away matches compared to other teams. But since we have low amount of data the CI's might not be meaningful. (Around 30 matches in away and home for each teams)

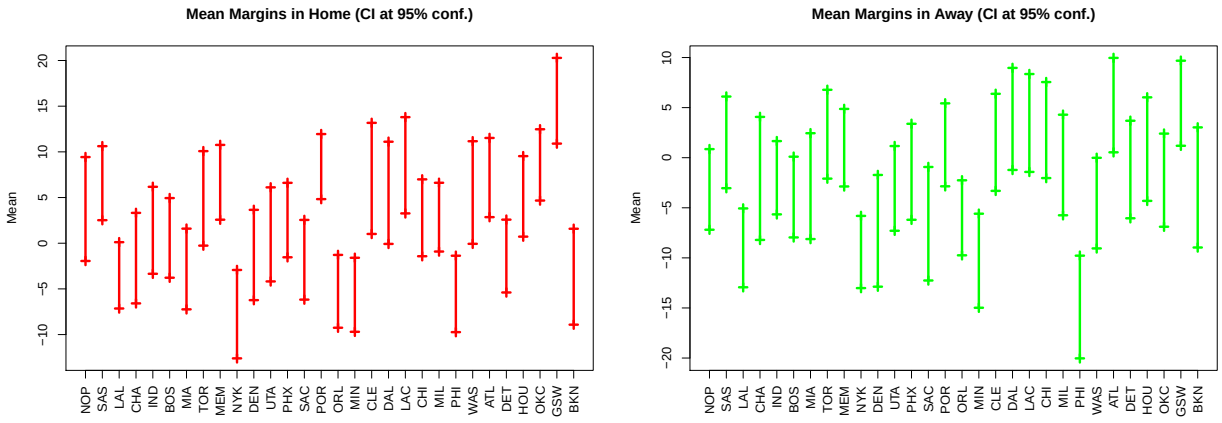

Figure 5: Confidence intervals of each nba team margins with respect to location of Home and Away.

In Figure 5, we observe that the Confidence interval of OKC as the closest to GSW, so we should compare their margin means to see if there is a significance.

```
## [1] F-test between GSW and OKC for Home p-val=0.4017
```

```
## [1] t-test between GSW and OKC for Home p-val=0.0214
```

```
## [1] Wilcox-test between GSW and OKC for Home p-val=0.0187
```

As it can be seen, even though, the p-value of the F-test is large and do not mentions about difference in margin variances for GSW and OKC, the p-values for the t-test and non-parametric wilcox-test(Anesthesiol., 2016) are particularly small and we obtain powerful evidences to state that the mean margin value of GSW is larger than OKC, and most likely GSW has the largest mean margin for the games played in home.

In Figure 5, we observe that the confidence interval of the PHI as the lowest, thus we will compare it with one of the teams with closer confidence interval. Thus we will compare PHI with DEN.

```
## [1] F-test between PHI and DEN for Home p-val=0.7863
```

```
## [1] t-test between PHI and DEN for Home p-val=0.0221
```

```
## [1] Wilcox-test between PHI and DEN for Home p-val=0.0142
```

As it can be seen above, we have a large p-value and we do not have any evidence to state any difference in margin variances in away location between these teams. Since we have low p-values for t-test and wilcox-test we have strong proofs to suggest that, PHI has lower margin means than DEN and most likely PHI has the lowest margin means compared to other teams.

2.2 The Effects of Game Locations

In the data given, when the matches take place in home stadium, the margin has higher average and median values compared to the matches played in away stadiums. We will apply some statistical tests to seek any interesting statistical evidences.

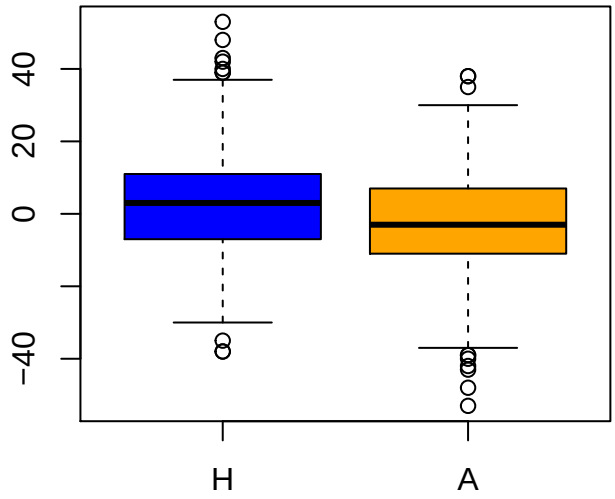

Figure 6: Comparison of margins in each match for Away and Home locations

Since we are comparing margin values, the variance will be equal to 0. Furthermore, the addition of each values in home and away margins will be equal to 0 as well. When we are looking at sport data like football or basketball, when we add all the margin values of each teams(played in home and away) we get 0. in Figure 6

```
## [1] F-test: p-val=1.0000
```

```
## [1] t-test: p-val=0.00000000001
```

```
## [1] Kruskal-test: p-val=0.0014
```

As the p-value's are less than the specified significance level 0.05 for both t-test and kruskal-test, we have evidence to conclude that the samples of the margin values for the matches played in home is larger than the games played in away stadiums.

When we apply prop test to examine the win rates of the teams for the different locations matches take place:

```
## [1] Prop-test: p-val=0.0000002416
```

Thus we have a solid proof to state that the win rates of the teams higher when they play the matches in home compared to the away.

3 Differences Between Periods

As the game approaches its end, we can consider that the performance of the players is affected and teams' approach to win the game might be also affected. In this section, to support our general thoughts(claims), we are going to investigate the 2 and 3 pointer attempts statistics and the connection of shoot results, with respect to periods. We will also look at the effects of location, period and the interaction of location and period over the 3 pointer shoot attempts via using two-way anova test(Endod., 2014). Furthermore, we are going to analyze the shoot distance preferences by each period.

3.1 2 and 3 Pointer Attempts Comparison Between 4 Main Periods

In the figure 7 we see the confidence intervals of the 2 and 3 pointer proportions to the total attempts of 2 and 3 respectively. The first noticeable things are the difference between the CI interval proportions of 2 pointers between period 1 and 4. Furthermore we also observe the difference of CI interval proportions of 3 pointers between period 1 and 4. Thus we are going start with analyzing these highlights.

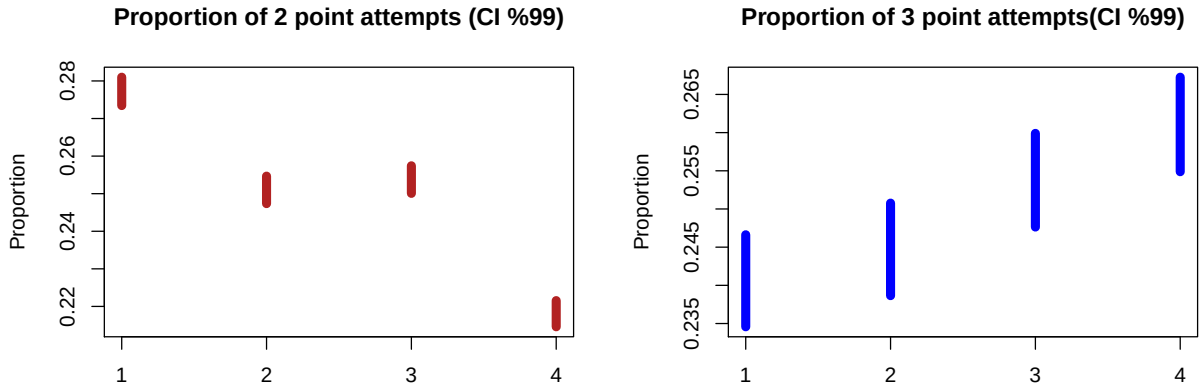

Figure 7: Confidence intervals for the ratios of 2 and 3 point attempts for each period to the total attempts made for all the periods

When we apply the prop test to the 2 pointer attempts between periods 1 and 4, we see huge difference between the proportions. We also obtain a very small p-value. So, we can say that, we have a beneficial evidence to state 2 pointer attempt proportion in period 1 is higher than period 4.

```
## $estimate
##    prop 1    prop 2
## 0.2772109 0.2180510
```

```
## [1] 5.22977e-193
```

Next, we look at the 3 pointer attempts between period 1 and 4 and obtain small p-value. So, we have a statistical demonstration to say the 3 pointer attempts is higher in period 4 compared to the period 1.

```
## $estimate
##    prop 1    prop 2
## 0.2405603 0.2610341

## [1] 5.013161e-10
```

- Perhaps, as the game gets closer to end, the teams start to play risky and focus on attempting 3 pointers rather than 2 pointers because they probably see the 3 pointers as quick and easy points. Thus we have some evidences to say the amount of 2 pointers probably decrease and 3 pointers likely to increase as the game ends.

3.2 Connection of Successful 2 and 3 Pointers with Periods

In the table below, we see the amount of shot-result by each main periods.

```
##             PERIOD
## SHOT_RESULT     1     2     3     4
##      made   12636 11451 11677  9881
##      missed 13253 11990 12020 10483
```

As we apply chi-test below, we do not have a small p-value thus we cannot mention about a strong relationship between shot result of 2 pointers and periods.

```
##
##  Pearson's Chi-squared test
##
## data:  period_table_2_points
## X-squared = 2.5949, df = 3, p-value = 0.4584
```

In the table below, we look at the shot result of 3 pointers with respect to periods.

```
##             PERIOD
## SHOT_RESULT    1    2    3    4
##      made   3004 2827 3048 2936
##      missed 5068 5383 5466 5823
```

In the chi-test below, we have a strong proof to say there is a significance between shot result of 3 pointers and the periods because we have a small p-value.

```
##
##  Pearson's Chi-squared test
##
## data:  period_table_3_points
## X-squared = 28.662, df = 3, p-value = 2.638e-06
```

3.3 Location vs Periods for 3 Points Shoot Attempts

We implement a two-way anova test to investigate how the mean of numerical variable differ according to 2 categorical variables. In our two-way anova test, we are going to look at how location, period and their combination affect the 3 pointer attempts.

As we only find a significance in the p-value of periods, we will conclude that we only have a evidence to suggest periods affect the amount of 3 pointers.

```
summary(two_way_anova)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## period              3    213   71.01  13.395 1.03e-08 ***
## location            1      7    6.93   1.307    0.253
## period:location     3      1    0.38   0.071    0.976
## Residuals        7100  37636    5.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4 Shoot Distances By Each Period

Figure 8 shows the distances of the shoot preferences in each period. As first sight, we see the period 4 has the highest mean of shoot distance according to data provided.

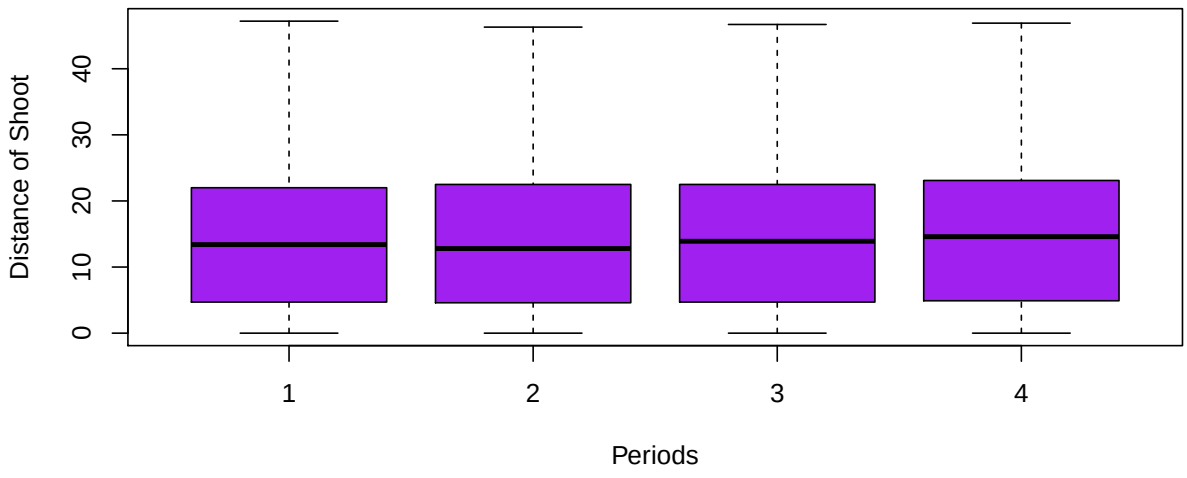

Figure 8: Shot Distances with Respect to Periods

In table 2, we applied 3 tests for homogeneity of variances at the significance level of 0.01 and by looking at their p-values, we can suggest that we have evidences to say variance of shoot distances differ across periods.

Table 2: Variance tests of the shot distances with respect to Periods

Test	p.value
Bartlett	0.00e+00
Fligner-Killeen	2.52e-05
Levene	0.00e+00

As we used one-way anova test below, we have a significant p-value and we have a proof to say there is a relationship between shot distances and periods.

```
##
##  One-way analysis of means
##
## data:  SHOT_DIST and PERIOD
## F = 39.441, num df = 3, denom df = 126942, p-value < 2.2e-16
```

- Previously, as we find evidences to state 3 pointer attempts in period 4 are likely to be highest among all periods, and this situation might cause an increase in the shoot distance preferences as well. Thus as we see in the table 2, the box-plots can be more meaningful to say as the match approaches to end, the 3 pointer attempts increase and as a result, the mean on shoot distance may increase.

4 The Effect of Shot Clock on 3 Pointers

As we see in many nba matches, the 3 point attempts in the last seconds of shot clock can be crucial for the result of the matches. Thus, analyzing the effects of shot clock can be considerable.

In the Figure 9 we look at the confidence interval of successful 3 pointers for the intervals of 0 to 2 seconds, 2 to 24 seconds and 0 to 24 seconds. As we can see, the confidence interval of the probability of successful 3 shoots in less than 2 seconds is lower than all samples and the 2 to 24 seconds interval. Since the sample size is also very small compared to other intervals, the 0-2 shot clock interval almost do not affect the shot clock probability confidence interval for the full sample size.

```
## [1] Prop-test: p-val=0.00000000000000014261
```

When we apply the prop-test we also see a very low p-value and since we obtain a strong proof to state that the probability of successful shoots is lower in the interval of 0-2 seconds than the interval of 2-24 seconds.

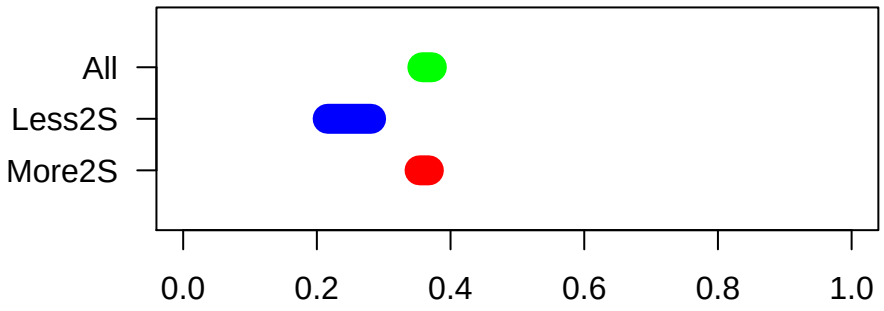

Figure 9: Comparison of probabilities for the successful 3 pointers with different shot clock intervals

5 Average Score Predictions of Nba Players

We calculate some average statistics for the players in given nba data set to perform a multi-linear regression among some specific features by looking at the partial correlation map (See the appendix table 3 to look at the partial correlation map created for the player analysis.) to predict the average score of the nba players.

After performing spearman correlation tests for the features dribbles_avg, touchtime_avg and point_3_success_percentage, we find $1.8041307 \times 10^{-34}$, 5.647874×10^{-65}, 4.3095135×10^{-6} respectively. As all of the features are significant at 0.01 confidence interval, we perform a multi-linear regression to predict the average score of the nba players via using these features.

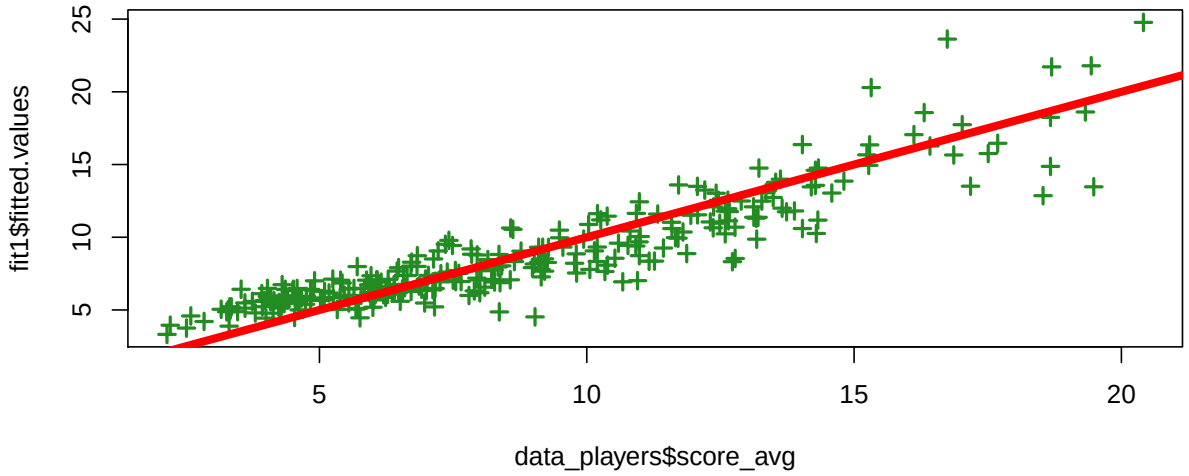

Figure 10: Multi-linear regression to predict the average score of the players

```
## [1] R_square_fit1 = 0.8270
```

```
## [1] Mean Absolute Error = 1.2936
```

```
## [1] Mean Absolute Percentage Error = 0.1777
```

We find 82.7% of the data fit the model, we make predictions with 1.2936 error for average score and our mean percentage error is 17.77%.

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.04815151       1.90295   0.444
##  Alternative hypothesis: rho != 0
```

Autocorrelation calculates the relationships of observations with different time periods. In both tests, we didn't see any significance, so we are failed to reject the null hypothesis and we have an evidence to conclude there is no autocorrelation on our residuals(differences between prediction and observation data).

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals[sample(1:sample_size, 200)]
## W = 0.96208, p-value = 3.342e-05
```

We have a p-value with less than 0.05 for the Shapiro-Wilk normality test, so we have a strong argument to say the residuals are not normally distributed.

```
##
##  studentized Breusch-Pagan test
##
## data:  fit1
## BP = 28.898, df = 3, p-value = 2.353e-06
```

Residual homoscedasticity checks whether residuals or forecasting errors have a constant variance. So the forecasting errors are heteroscedastic since our alpha is greater than p-value and we will conclude the residuals do not have constant variance.

```
outlierTest(fit1)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 33 -4.383216         1.6638e-05    0.0046753
```

```
data_players$player_name[33]
```

```
## [1] "Kobe Bryant"
```

In our fitted model, we only find Kobe Bryant as the the outlier, the person with score average that differs significantly from other players.

6 Conclusion

In the report, we examined the features of NBA dataset in detail. We calculated some margin mean, median and variance for the NBA teams with respect to home and away locations. We examine the average scores of the players in a histogram with the average confidence interval included(The average do not include free throws). Furthermore we investigate some key features with correlation heatmap. When it comes to the inferential statistics, we take advantage of various statistical techniques to make inferences from the data. We take advantage of some tests such as t-test,F-test, Chi-Test, One Way Anova, Two Way Anova. We also performed some non parametric tests to compare the variance of the team margins. In addition, we computed and visualized a linear regression for the prediction of average point scores and we evaluated the performance of our linear regression deeply.

When it comes to summarize our key points:

1.There is a strong correlation between some particular player average features. (Section 1)

2.Match locations impact the margin performance of the teams significantly. (Section 2)

3.Match locations also greatly affect the odds of teams to win the match. (Section 2)

4.Periods have major impact on the amount of attempts for 2 and 3 pointers. (Section 3)

5.There is a strong relationship between the probability of successful 2 and 3 pointers and the periods. (Section 3)

6.There is a connection between shoot distances and the periods. As the match approaches to the end, players tend to prefer 3 pointers and that might be the consequence of the increase in shoot distance average. (Section 3)

7.As the shot clock approaches to end, successful 3 point shoot rates reduce. (Section 4)

8.Average score predictions can be made via using particular features of the players. (Section 5)

These observations appear statistically significant and further analysis can be made to check our findings.

References

Anesthesiol., K. J. (2016). Nonparametric statistical tests for the continuous data: the basic concept and the practical use. page 1.

Endod., R. D. (2014). Statistical notes for clinical researchers: Two-way analysis of variance (anova)-exploring possible interaction between factors. page 2.