A study of the 2014-2015 NBA dataset

Harout Barikian

08 February 2021

Abstract

We displayed how we can apply diverse descriptive and inferential statistical tools to the NBA dataset. Teams playing at home might have a higher chance of winning than teams playing away. Also, teams playing with a three points style might be more efficient to win games; there was a strong positive correlation between the three points per game scored and the win rate of a team. Then we saw that the points per game scored by players between different teams were homogeneous for the variances. Besides, the median of points per game scored by the players was not likely impacted by the rank of their respective teams. In the end, we analysed the accuracy of each shot taken compared to the distance and it appears to be a strong negative correlation between the two variables. But for Kyle Korver, the distance seems to not have an impact on his shooting accuracy.

Contents

1 Introduction

The National Basketball Association, believed by many to be the best basketball league in the world. In this report, we are going to analyse the 2014-2015 regular season. In the first part, the focus will be on overall Team performance and in the following part, we will be talking about player-related performance.

To begin with, the home and away win percentage will be tested to see if teams usually win more when playing at home instead of away. Furthermore, we will try to see the playing style of each team, meaning if a team is revolving around shooting more three points and playing near the arc or go more inside near the basket and score two points. After that, we will check if there is any correlation between the number of three points per game scored and the win percentage of a team. In the end, we will study the points per game scored by each player in a team and look if there is any significant dispersion between the points. We will also analyse if there is an impact on the points per game by each team's ranking.

Going into the second part, we will test to see if the shooting distance affects the scoring ability of all the players in the league. Then we will rank all the players by their three points per game scoring average, picking $Kyle\ Korver$ and then analysing his scoring frequency depending on the shooting distance. In this way, we can try to understand if the top player is also affected or not by the distance factor like all the other players.

2 Background

To establish the ideas and topics that we will talk about in this report, many articles were found online that inspired and may justify the hypothesis presented. According to (Entine, 2007), it seems that Home court advantages were related to the resting period teams were having as the away team was less rested due to going on the road. In addition, according to (Cheema, 2021), referee bias was found for the home team as they would call for fouls more often against the away team. All of this seems to support and explain the results we got.

(Babb, 2013) mentions in his article how the introduction of the three points in a team's play helps in having more options in the offense, like opening lanes for penetration and creating room for operation under in the post. He also talks about $Kyle\ Korver$ and how good of a shooter he is, describing him with 'This guy's shot always looks like it's going in'. In another article written by (Goldsberry, 2019), he talks about the effect of three points scored on the team's win conditions, with Popovich(former San Antonio Spurs coach) stating 'Now you look at a stat sheet after a game and the first thing you look at is the 3s. If you made 3s and the other team didn't, you win'. The NBA seems to be revolving around the three points mark after all.

Lastly, we will check the disparity in points per game between players in a team, as (Grant, 2018) said, having too many star players in a team affected the performance because everyone wanted to be the alpha dog. Teams with only three star players won more games than teams with four or five.

After all the interesting ideas are read and seen, we will dive right into the statistical analysis of this NBA dataset.

Table 1: First five rows of the Home and Away win percentages per Team

	TEAM	H_WIN_PCT	A_WIN_PCT	DIFF_PCT
28	GSW	93	69	24
30	ATL	87	71	16
15	POR	81	52	29
9	MEM	79	69	10
2	SAS	78	50	28

3 Team-based hypothesis

3.1 Home/Away win percentage comparison

Teams in sports competitions tend to win more playing home games rather than away. Table 1 shows some of the teams home and away win percentages, with the column (DIFF_PCT) calculating the difference between the two values. For example, Portland Trail Blazers POR seems to struggle to win away with a 29% deficit when not playing at home. We are going to test to determine whether the true mean of home wins is larger than the true mean of away wins for all the teams.

```
two sample t-test                Multivariate Shapiro-Wilk normality test
p-value: 0.0043041               w=  0.9510474 ; p-value 0.1803297
confidence interval: 4.809622 Inf
sample estimates: 56.3 43.83333
```

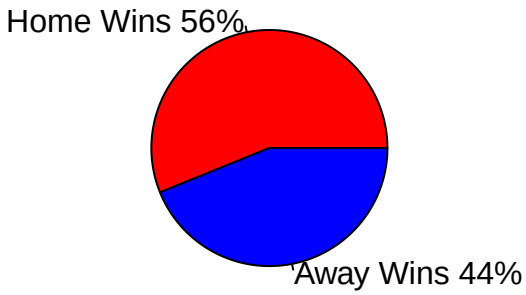

Figure 1: Home and Away win percentages

Even though the number of observations $n = 30$ is not too large, there are some pieces of evidence suggesting that the data follow a gaussian distribution with a p-value of 0.18 for the multivariate shapiro-wilk test. Therefore we do not reject H_0 for H_1 at the significance level $\alpha = 0.05$, making the Gaussian approximation to be relatively accurate.

The obtained p-value for the two sample t-test of 0.004 is notably smaller than the considered significance level $\alpha = 0.05$. We thus reject H_0 for H_1 at the significance level $\alpha = 0.05$. There are strong statistical evidences suggesting that the teams playing at home might have a higher chance of winning than the teams playing away, as shown in figure 1.

3.2 Attacking team playing style

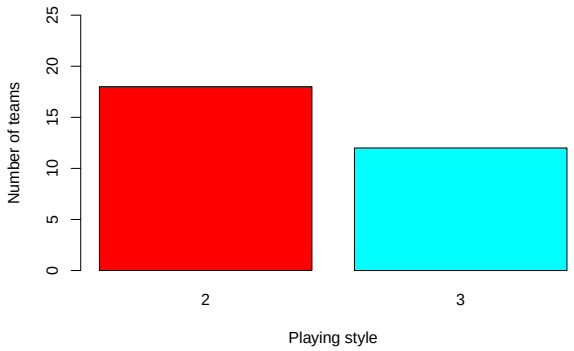

Figure 2: Barplot for the number of teams for each playing style

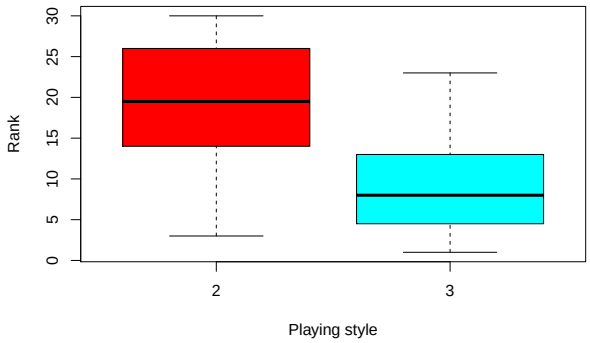

Figure 3: Boxplot for the teams ranking and their playing style

To qualify a team into a specific category, we assumed that if a team scores three points per game more than the mean of all teams, it would be categorised as a 'three points style of play'. Else, it would be in the 'two points style of play'. The mean is equal to 19.16 three points per game, rounded to a 19 to make it more practical.

In the barplot shown in figure 2, we can clearly notice that the teams playing a three points style of game (12) are less than those playing more of a two points style of game (18); this could be because it is a more difficult play style as you need very good distance shooters in the team.

In the boxplot represented in figure 3, we can see that the teams using a three point style look to be better in the rankings compared to the opposite two point style of play. We will test this hypothesis by comparing the true mean of win percentage from each style of play. After that, we will check if there is a correlation between the win percentage of a team and the three points per game scored.

```
two sample t-test                    Shapiro-Wilk normality test

p-value: 0.0004083748                 w=  0.9629931 ; p-value 0.3685581

confidence interval: 10.61389 Inf

sample estimates: 61.6025 42.27
```

Even though the number of observations $n = 30$ is not too large, there is strong evidences that the data follow a gaussian distribution with a p-value of 0.36 for the shapiro-wilk test. Therefore we do not reject H_0 for H_1, making the Gaussian approximation to be relatively accurate.

The obtained p-value for the two sample t-test of 0.0004 is very small, smaller than the considered significance level $\alpha = 0.05$. We thus reject H_0 for H_1 at the significance level $\alpha = 0.05$, these are strong statistical evidences suggesting that the teams with a three point playing style are more likely to have a higher win percentage than the teams focusing more on a two points style of play.

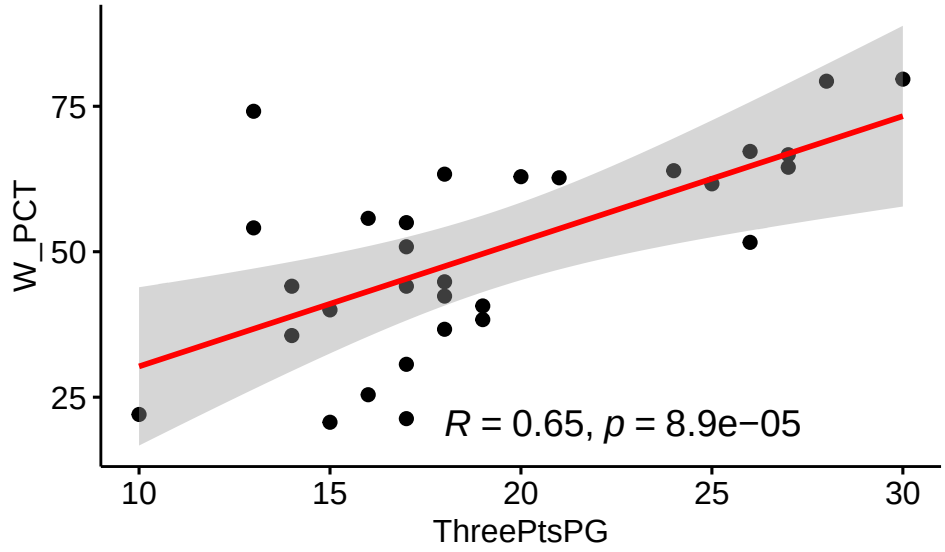

Figure 4: Scatter plot and correlation results between three points per game and win percentages

```
Multivariate Shapiro-Wilk normality test

w=  0.9414386 ; p-value 0.09944993
```

The data in figure 4 suggest that there is a statistically significant positive correlation between the team's win percentage and their three points per game score.

At a significance level of $\alpha = 0.05$ we observe a correlation of $R = 0.65$ with a very small p-value. The multivariate normality assumption does seem to be verified with a p-value of 0.099 which is marginally larger than the significance level $\alpha = 0.05$; The result of our correlation analysis might be reliable. Thus there would possibly be a correlation between the three points per game scored and the win rate of a team.

3.3 In team's points dispersion

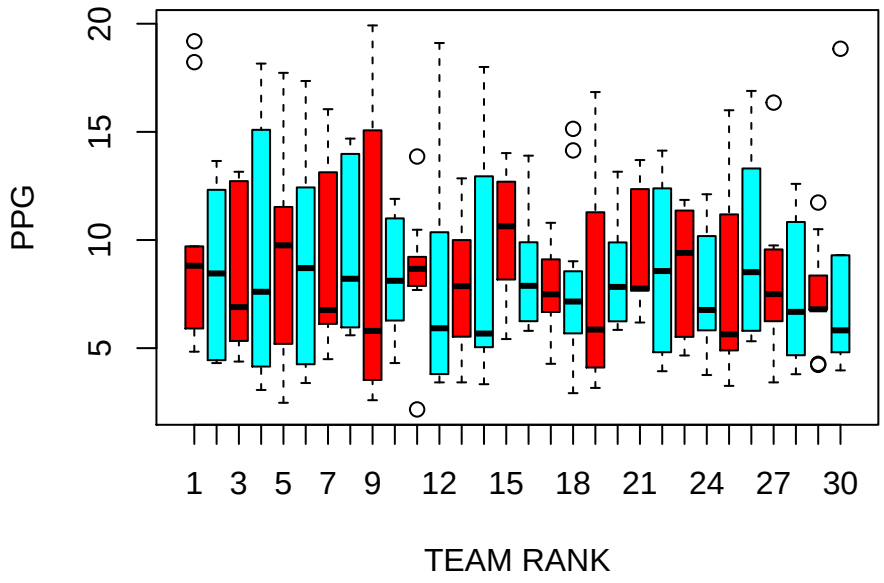

Figure 5: Boxplot of each player's points per game by its team rank

In figure 5, we can see the points per game scored by each team and their placement in the ranking. We can notice some outliers in this boxplot, the number one ranked team has two outliers with high scoring average, which shows that this team have already two of the very best players, much better than the other players from that same team. on the other hand, we can also see an outlier in one of the lowest ranked teams.

To examine more the outliers we just talked about, we will use the a pie chart of the points per game scored by each player in the team. We will look for the first ranked team GSW $Golden\ State\ Warriors$ and the lowest ranked team NYK $New\ York\ Knicks$.

From the two pie charts shown in figure 6, we can clearly see that $Stephen\ Curry$ and $Clay\ Thompson$ have very high points per game compared to their teammates, with $19.19PPG$ and $18.22PPG$ each. In addition, $Carmelo\ Anthony$ is the top scorer in his team with a $18.84PPG$; $9PPG$ more than the second-best in the team ($Time\ Hardaway\ Jr$ with $9.3PPG$).

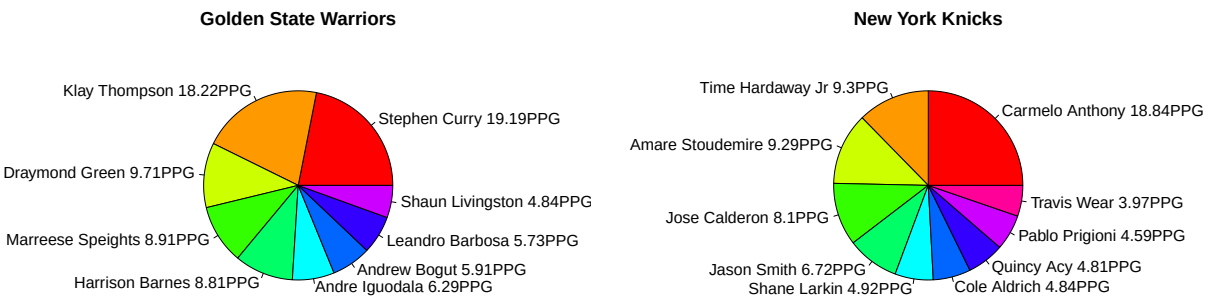

Figure 6: Pie charts for the points per game scored by the Golden State Warriors and New York Knicks players

We are going to test the equality (homogeneity) of variance across the teams, using the fligner, bartlett and Levene tests.

```
    Fligner-Killeen test of homogeneity of variances

data:  PPG by TEAM_RANK
Fligner-Killeen:med chi-squared = 27.853, df = 29, p-value = 0.5258
```

```
    Bartlett test of homogeneity of variances

data:  PPG by TEAM_RANK
Bartlett's K-squared = 37.668, df = 29, p-value = 0.1299
```

```
Levene test of homogeneity of variances
```

	Df	F value	Pr(>F)
group	29	0.9719712	0.5113395
	248	NA	NA

From the output of the fligner, bartlett and Levene tests, we see that the p-values are relatively large ($0.52, 0.12$ and 0.51). This means that there is no statistical evidence that the variances are significantly different (for instance, at $\alpha = 0.05$, we do not reject the null hypothesis). We can therefore assume there is a homogeneity of variances in the different groups.

```
    Shapiro-Wilk normality test

data:  res.aov$residuals
W = 0.95366, p-value = 1.011e-07
```

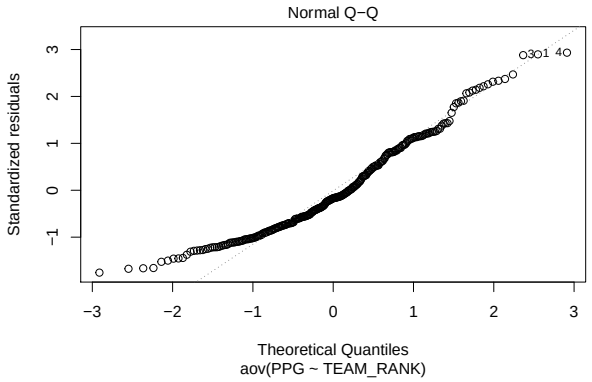

Figure 7: Qunatile-quantile plot of the anova residuals against the expected normal distribution

The very small p-value of the Shapiro-Wilk normality test and the qqplot shown in figure 7 suggests that we should reject H_0 for H_1. There are statistical evidences suggesting that the residuals may not be Gaussian. Because of this, we cannot use the $anova$ test as the conditions are not met. We will therefore use the $Kruskal - Wallis$ test that is an alternative for $anova$.

```
    Kruskal-Wallis rank sum test

data:  PPG by TEAM_RANK
Kruskal-Wallis chi-squared = 11.591, df = 29, p-value = 0.9983
```

From the $Kruskal - Wallis$ test just performed, the obtained p-value is very high (0.99); for instance we do not reject H_0 for H_1. The test thus suggests that the median of PPG scored by the players are not impacted by the rank of that team.

4 Player based hypothesis

4.1 General shooting accuracy

In the histogram from figure 8, we can see in the histogram how many shots have been taken from each distance. Around $3feet$ and $25feet$, we can notice the most number of shooting attempts from the players. In the boxplot from figure 8, shots made and missed are being compared to the distance using a boxplot. We can visually assume that the mean distance of shots made($10feet$) is smaller than the mean distance of shots missed($18feet$). This is logical because shooting closer to the basket should usually result in a higher chance of scoring.

The entry of the dataset for the distances does not really consist of "the realisation of a random variable". Therefore it is more meaningful to consider the $Spearman$ correlation instead of the $pearson$ correlation. Thus the data indicate a very strong negative correlation ($R = -0.82$) with a very small $p - value$ between Distance and Accuracy, as shown in figure 9; the variable Accuracy seems to decrease almost linearly with the Distance.

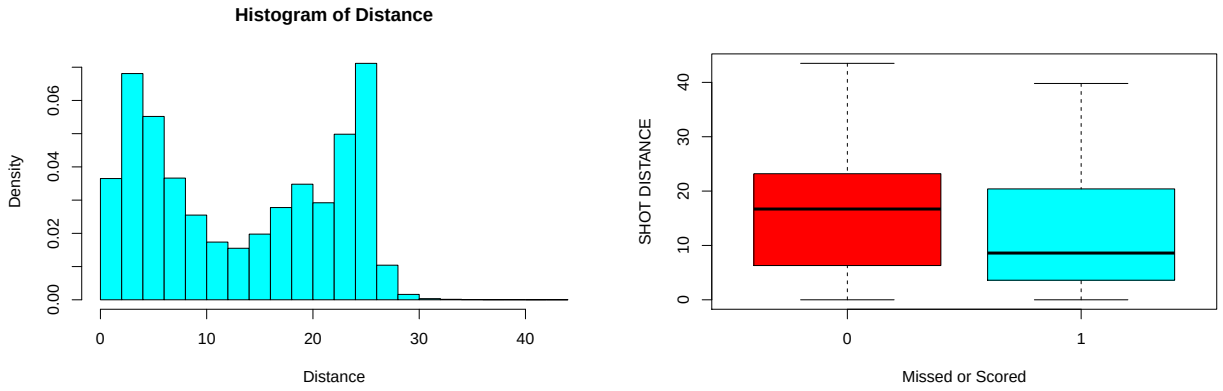

Figure 8: Histogram of the shot's distances density and a Boxplot of the shot's result from the distance taken

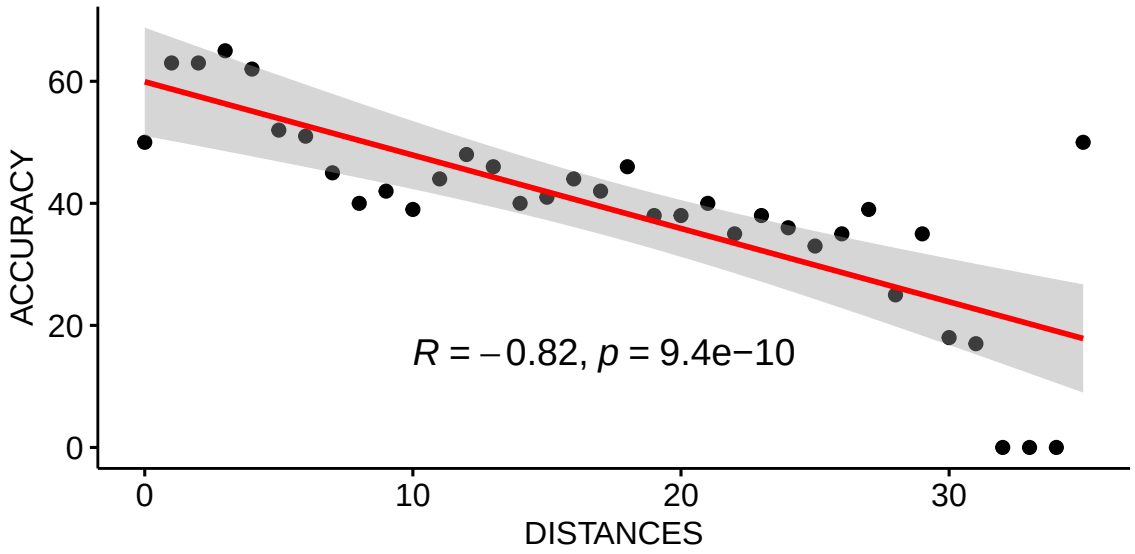

Figure 9: Scatter plot and correlation results between the distance and the shot's accuracy

Table 2: The best three point shooters

	P_NAME	ACC_THREE	THREE_ATTEMPTS
115	Kyle Korver	49	336
198	Klay Thompson	44	395
239	Jj Redick	43	322
197	Stephen Curry	42	436
220	Wesley Matthews	40	419

Therefore we might suggest that shooting from a further distance negatively impact the scoring chances, shooting from a closer range gives more successful results. Since the sample size is $n = 36$, the result of our correlation analysis might be reliable.

4.2 Kyle Korver's shooting accuracy

In table 2, we can check the top three point scorers in the 2014-2015 NBA regular season. $Kyle\ Korver$ is the best three point shooter with a 49% accuracy, 5 more than the runner-up $Klay\ Thompson$. We can notice that $Kyle\ Korver$ has 336 attempts, much lower than $Stephen\ Curry$ with 436 shots taken. This might be because of the number of games played, injuries, and a lot of other factors that can affect these values.

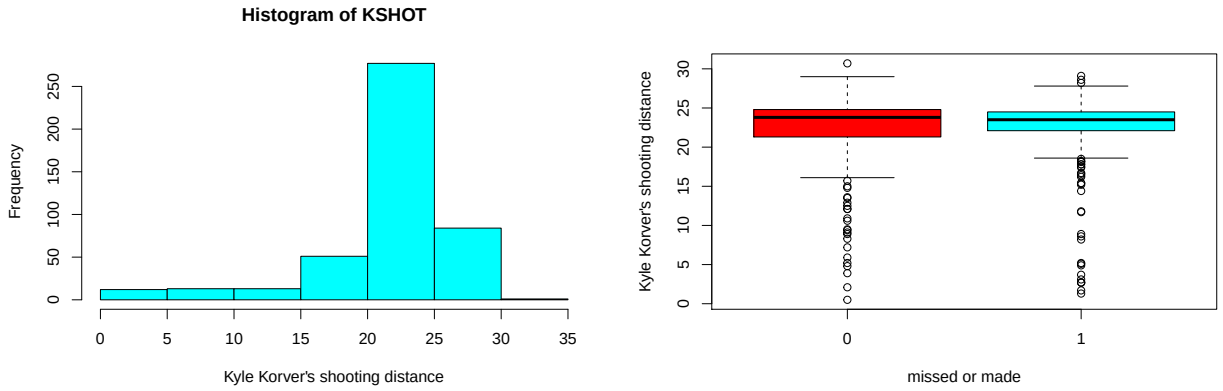

Figure 10: Histogram of Kyle Korver's shot's distances density and a Boxplot of his shot's result from the distance taken

In the histogram from figure 10, we see the number of shots taken by $Kyle\ Korver$ from each distance, we realise that his favorite distance to shoot from is approximately between 20 and 25 feet. In the boxplot from the same figure, we compare the shots made or missed with each distance taken from. There does not seem to be any relevant difference in the scoring between the distances he is shooting from. We will test this hypothesis with a t-test between shots taken before and after $22 feet$. We chose this distance because it is the start of the three point line.

```
two sample t-test

p-value: 0.4951784

confidence interval: -0.1448356 0.07029915

sample estimates: 0.4553571 0.4926254
```

The obtained p-value is very large $p - value = 0.49$, so that we do not reject H_0 for H_1 and we thus conclude that the number of points scored from a distance more or less than 22 are likely the same for $Kyle\ Korver$. The sample size is very large, so that the previous conclusion can be reliable. We may find this observation to be usually not what would we expect as players would score more from a closer distance to the basket, but in this case, $Kyle\ Korver$ likes to shoot from long distances, he has 339 attempts from beyond the arc of three point with only 112 attempts from inside the line. In addition, he is the top three point scorer in the league, this can be an explanation as to why there is no difference between the points scored from distances for him.

5 Conclusion

To conclude this report, we used various descriptive and inferential statistics to analyse the NBA 2014-2015 regular season dataset. We performed $t-tests$ for the means between two samples, some parametric($pearson$) and non-parametric($spearman$) $correlation$ analyses, also a $Kruskal-Wallis$ test for the median of multiple samples which was an alternative to the $anova$ test. In addition, we checked all the conditions required for each parametric test with some $shapiro-wilk$ test for normality if the sample size was not big enough. Furthermore, we used $Fligner-Killeen$, $bartlett$ and $Levene$ tests to check for homogeneity of variances. Moreover, many figures were used to describe and visualise the data, using $histograms$, $plots$, $pie-charts$ and $boxplots$ wherever necessary.

To sum up the main ideas that we can extract from this report:

1. Teams playing at home might have a higher chance of winning than the teams playing away (Section 3.1);
2. Teams with a three point playing style are more likely to have a better ranking than the teams focusing more on a two points style of play (Section 3.2);
3. A possible correlation between the three points per game scored and the win rate of a team (Section 3.2;
4. Homogeneity of variances(points per game scored by players) between the different Teams (Section 3.3);
5. The median of points per game scored by the players might not be impacted by the rank of their respective Team (Section 3.3);
6. Shooting from a further distance might negatively impact the scoring chances, shooting from a closer range can give more successful results (Section 4.1);
7. The shooting distance for $Kyle\ Korver$ does not seem to affect his shooting accuracy (Section 4.2).

The results shown appear to be statistically significant and relevant. Nevertheless, to be even more

accurate and precise in our analysis, we could use more datasets from other years of the NBA season's to strengthen all the observations and deductions made.

6 References

1. Entine, O.(2007). *The Role of Rest in the NBA Home-Court Advantage* Available at: http://www-stat.wharton.upenn.edu/~dsmall/nba_rest_submitted.pdf
2. Cheema, A.(2021). *Does NBA Officiating Favor the Home Team?* Available at: https://www.thespax.com/nba/does-nba-officiating-favor-the-home-team/
3. Babb, S.(2013). *How the 3-Point Shot Has Revolutionized the NBA* Available at: https://bleacherreport.com/articles/1715367-how-the-3-point-shot-has-revolutionized-the-nba
4. Goldsberry, K.(2019). *The NBA is obsessed with 3s, so let's finally fix the thing* Available at: https://www.espn.co.uk/nba/story/_/id/26633540/the-nba-obsessed-3s-let-fix-thing
5. Grant, A.(2018). *The Problem with All-Stars* Available at: https://www.linkedin.com/pulse/problem-all-stars-adam-grant

7 Appendix

Table 3: Team's playing style and win percentage

RANK	TEAMS	TwoPtsPG	ThreePtsPG	STYLE	W_PCT
1	GSW	53	30	3	79.66
2	ATL	51	28	3	79.31
3	MEM	58	13	2	74.14
4	POR	43	26	3	67.24
5	HOU	35	27	3	66.67
6	LAC	55	27	3	64.52
7	DAL	50	24	3	63.93
8	CHI	53	18	2	63.33
9	CLE	45	20	3	62.90
10	SAS	49	21	3	62.71
11	TOR	48	25	3	61.67
12	OKC	46	16	2	55.74
13	WAS	61	17	2	55.00
14	NOP	52	13	2	54.10
15	PHX	52	26	3	51.61
16	MIL	45	17	2	50.85
17	IND	51	18	2	44.83
18	CHA	57	14	2	44.07
19	MIA	45	17	2	44.07
20	BKN	53	18	2	42.37
21	BOS	52	19	3	40.68
22	UTA	50	15	2	40.00
23	DET	44	19	3	38.33
24	DEN	51	18	2	36.67
25	SAC	58	14	2	35.59
26	ORL	52	17	2	30.65
27	LAL	50	16	2	25.42
28	MIN	48	10	2	22.03
29	PHI	37	17	2	21.31
30	NYK	41	15	2	20.69

Table 4: Accuracy(in percentage) for each shooting distance(in feet) for all players

DISTANCES	ACCURACY	SHOTS_TAKEN
0	50	4
1	63	420
2	63	800
3	65	893
4	62	754
5	52	669
6	51	497
7	45	472
8	40	382
9	42	340
10	39	252
11	44	198
12	48	172
13	46	193
14	40	200
15	41	241
16	44	300
17	42	339
18	46	382
19	38	436
20	38	450
21	40	330
22	35	405
23	38	593
24	36	983