Basketball: The Trend of Point Guard's Offensive Abilities which are Shooting, Steal, Assist and Turnover

Yingxuan Shi

26/April/2022

Abstract

The point guard, as a core and soul of a basketball team, plays an important role in offense. This paper utilizes data from NBA stat website: analyzing point guards' offensive abilities which are shooting, passing, steal, assist and scoring. James Harden, Kyrie Irving and Stephen Curry are main observations in the paper, James Harden is good at assist, free throw and turnover; Kyrie Irving had the highest Plus Minus in them; Stephen Curry has prominent field goal percentage, 3-point field goal percentage and steal. Therefore, people can find the future trend of being a good point guard through reading the paper.

Contents

1. Introduction **2**

2. Data Section **2**

2.1 Data Cleaning 2

2.2 Variables 3

3. Model Section **5**

3.1 Model 5

3.2 Regression Model Diagnostics 5

4. Result Section **6**

4.1 James Harden 6

4.1.1 Oklahoma City 6

4.1.2 Houston 8

4.1.3 Brooklyn 10

4.2 Kyrie Irving 12

4.2.1 Cleveland 12

4.2.2 Boston 14

4.2.3 Brooklyn 16

4.3 Stephen Curry 18

4.3.1 Golden State 18

4.4 Comparison 20

5. Discussion **22**

5.1 James Harden 22

5.2 Kyrie Irving 22

5.3 Stephen Curry 22

5.4 Limitations and Future 23

Apendix **24**

A Datasheet 24

1. Introduction

A good basketball team needs five players, each filling a specific role. Therefore, we have five positions: point guard, shooting guard, small forward, power forward and center (Chris 2020). Point guard is theme of my analysis. The point guard plays an important role in the team's offense so point guards not only need the ability to drive the lane, but also need a stable jump shot and 3-pointer. Moreover, point guards must be good at passing and ball-handling skills. On defense, they should find opportunities of steal and turnover (Chris 2020). In the paper, I will analyze three top point guards in NBA. They are James Harden, Kyrie Irving and Stephen Curry. The National Basketball Association (NBA) is a professional basketball league. It represents the highest level of basketball game and have the best basketball players in the world (Global 2021). Thus, I think analyzing NBA basketball players is most representative. Although James Harden, Kyrie Irving and Stephen Curry have same position (point guard), they have different styles of point guard. James Harden is good at generating fouls, free throws and step-back shots; Kyrie Irving has a strong ability to drive the lane; Stephen Curry is one of most famous 3-point shooters.

Analyzing point guard's field goal, 3 point field goal, shooting average, assist, steal, turnover is very important. This is because point guards are the heart and soul of their respective teams. They just like the brain of the team and decide the trend, speed, and victory of games (Global 2021). Each ability of point guards can affect their own performance, even the whole team. Furthermore, this analysis can be used in how to cultivate a good point guard by predicting the future trend of being a good point guard. I compare the players' strengths and weaknesses and analyze whether they affect their Plus Minus in the paper. James Harden and Stephen Curry are the future trend of being a good point guard. James Harden has prominent free throw percentage, assist and turnover and Stephen Curry has prominent field goal percentage, 3-point field goal percentage and steal.

This paper is organized as follows: In the Data Section, I introduce the raw data and select the variables that we will use in the paper. Then, I describe these variables using figures and tables. In the Model Section, I explain linear regression model appropriately and describe how to use it in my analysis. In the Results Section, I find field goals attempt, field goals made, 3 point field goal attempt, 3 point field goal made of James Harden, Kyrie Irving and Stephen Curry whether they correlate their Plus Minus. Moreover, I compare their offensive abilities, which are steal, assist, percentage of shots and scoring, using two radar charts and a box plot. In the Discussion Section, I discuss whether my results match the real cases, and then think about about limitation and future,

2. Data Section

2.1 Data Cleaning

To gain a better understanding of my favorite basketball players who are James Harden, Kyrie Iring and Stephen, I utilized 2004-2022 NBA games details from NBA stats website. The data is collected and provided by NBA games data (NBA Official). In this dataset, the raw data includes 29 variables so we cleaned and extracted the important data to start my analysis. In the analysis, I will use R statistical language (R Core Team 2019), tidyverse packages(Wickham et al. 2019), devtools (Wickham et al. 2021), dplyr (Wickham et al. 2022), fmsb(Nakazawa 2022), janitor (Firke 2021), formattable (Ren and Russell 2021), kableExtra (Zhu 2022).

Firstly, I filtered observations of James Harden, Kyrie Irving, Stephen Curry respectively because I will analyze them in the paper. For variables, I just selected PLAYER_NAME, TEAM_CITY, PTS, AST, FG3M, FG3A, FG3_PCT, FGA, FGM, FG_PCT, FT_PCT, STL, TO, Plus-Minus because I will analyze players' offensive aspect in the paper. PTS is player's score in a game; AST is assists; FG3M is 3 Point field goals made; FG3A is 3 Point field goals attempted; FG3_PCT is 3 point field goal percentage; FGM is field goals made; FGA is field goals attempted; FG_PCT is field goal percentage; FT_PCT is free throw percentage; STL is steal; TO is turnover; Plus_Minus reflects how the team did while that player is on the court. For James Harden, he just played 6 games in Philadelphia so I deleted them. Because I think these 6 observations do not need to be analyzed. Then, the new datasets still have several NA values because of some mistakes and absences so I deleted all NAs. Lastly, I divided datasets into several datasets to analyze players' data in different teams. For example, James Harden played in Oklahoma City ,Houston and Brooklyn.

2.2 Variables

To better understand the variables in my paper, I reported three bar plots using ggplot2(Wickham (2016)) and three tables using kableExtra(**p?**) to explain variables. Obviously, PLAYER_NAME only includes James Harden, Kyrie Irving and Stephen Curry. From figure1, figure2 and figure3, we can see the number of games in different TEAM_CITY. For James Harden, he played the most number of games in Houston (741 games). For Kyrie Irving, he played the most number of games in Cleveland (439 games). For Stephen Curry, he only played in Golden State (980 games). Table1, table2 and table3 show the rest of variables, PTS, AST, FG3M, FG3A, FG3_PCT, FGM, FGA, FG_PCT, Plus_Minus. These variables describe players' offensive data on the game. Numbers in red are 3 point field goal percentage and field goal percentage for each player, I think they can reflect players' offensive efficiency. Moreover, numbers in blue are Plus_Minus which reflects how the team did while that player is on the court. Plus_Minus is the most important variable in the data analysis. If Plus_Minus is positive, it means the player is good in the game. If Plus_Minus is negative, it means the player is not good in the game.

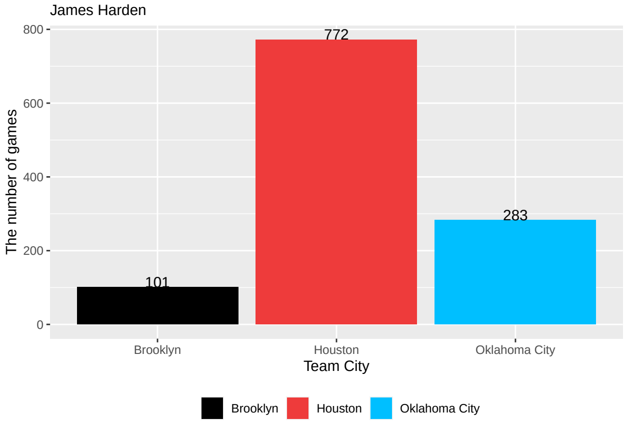

Figure 1: The number of James Harden's games in Houston, Oklahoma city and Brooklyn

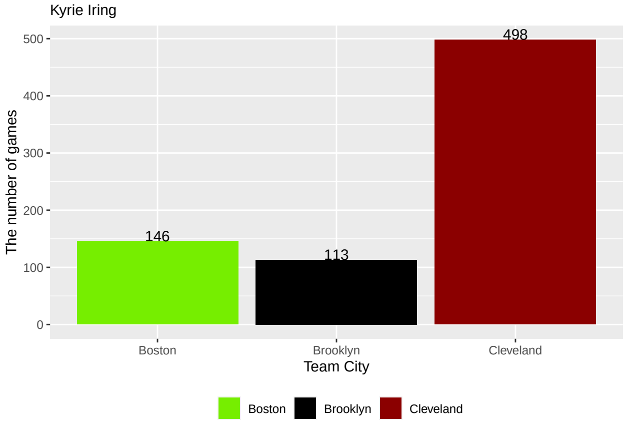

Figure 2: The number of Kyrie Irving's games in Brooklyn, Boston and Cleveland

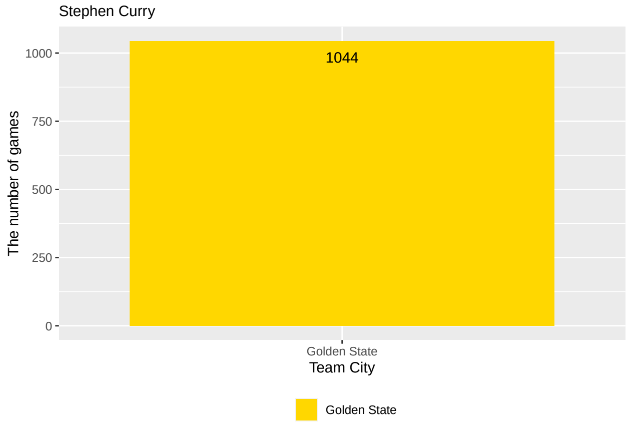

Figure 3: The number of Stephen Curry's games in Golden State

Table 1: James Harden's first 5 rows of cleaned dataset

PTS	AST	STL	TO	FG3M	FG3A	FG3_PCT	FGM	FGA	FG_PCT	FT_PCT	Plus_Minus
11	5	0	6	3	7	42.9	3	17	17.6	0.0	-30
16	14	1	4	1	5	20.0	5	15	33.3	77.8	12
25	11	2	6	2	4	50.0	6	10	60.0	100.0	6
26	9	1	6	2	6	33.3	8	13	61.5	100.0	9
29	16	0	3	3	7	42.9	8	14	57.1	100.0	19

Table 2: Kyrie Iring's first 5 rows of cleaned dataset

PTS	AST	STL	TO	FG3M	FG3A	FG3_PCT	FGM	FGA	FG_PCT	FT_PCT	Plus_Minus
22	5	1	1	5	11	45.5	8	17	47.1	50.0	27
50	6	1	3	9	12	75.0	15	19	78.9	84.6	14
19	6	3	3	2	6	33.3	8	18	44.4	100.0	-5
38	5	2	2	2	6	33.3	14	26	53.8	88.9	4
29	5	2	2	3	8	37.5	10	22	45.5	100.0	-5

Table 3: Stephen Curry's first 5 rows of cleaned dataset

PTS	AST	STL	TO	FG3M	FG3A	FG3_PCT	FGM	FGA	FG_PCT	FT_PCT	Plus_Minus
8	8	1	4	2	4	50.0	3	7	42.9	0	14
34	3	2	2	5	12	41.7	11	21	52.4	100	16
15	5	3	4	1	6	16.7	5	12	41.7	0	6
30	1	0	4	4	9	44.4	13	22	59.1	100	-12
21	9	1	3	4	5	80.0	8	15	53.3	100	-8

3. Model Section

3.1 Model

In the paper, I focus on relationship between Plus Minus and FGA, FGM, FG3M, FG3A so I will use the linear regression model. Linear regression can predict the value of an outcome variable (Y) by one or more input variables (X). It establishes a linear relationship between response variable (Y) and independent variable (X). Therefore, I can estimate the value of Y, when Xs are fixed. This is my model:

$$Y = \beta_o + \beta_1 X_{FGA} + \beta_2 X_{FGM} + \beta_3 X_{FG3A} + \beta_4 X_{FG3M} + \varepsilon$$

In the mathematical formula, Y is the outcome of the value of Plus Minus. β_o is the intercept, the predicted value of Y when the X is 0. β_1, β_2, β_3, $\beta_4$are the coefficients that Y increases or decreases as Xs increase. X_{FGA}, X_{FGM}, X_{FG3A}, X_{FG3M} are the independent variables (field goal attempt; field goal made; 3 point field goal attempt; 3 point field goal made). ε is variation there is in our estimate of the regression coefficient.

3.2 Regression Model Diagnostics

- Residuals vs Fitted plot: Check the linear relationship assumptions. A horizontal line, which have no distinct patterns, shows a good linear relationship.
- QQ plot: Examine whether the residuals are normally distributed. Residuals points following the straight line is good.
- Scale-location plot: Check the homogeneity of variance of the residuals. Points distributing equally around the horizontal line is good.
- Residuals vs Leverage plot: Identify influential cases that extreme values influence the regression results.

4. Result Section

4.1 James Harden

4.1.1 Oklahoma City

During James Harden's time in Oklahoma City, he as a rookie showed great offensive ability. From figure 4, we can see the linear relationship between Plus Minus, which is a statistics to evaluate a player's performance in the game, and scoring abilities, which are FGM, FGA, FG3M, FG3A. Field Goal Made (FGM) shows a strong linearly increasing relationship and its slope is higher than the others significantly. It means that FGM could affect significantly Plus Minus of James Harden in Oklahoma City. Then, I built a linear model to analyze the correlations between Plus Minus and FGM, FGA, FG3M, FG3A. We can see regression model diagnostics from figure 5. Residuals vs Fitted plot: there is no pattern in the residual plot so we can assume linear relationship between the predictors and the outcome variables; QQ plot: residuals points follow the straight line so the model is good; Scale-location plot: a horizontal line with equally spread points so the model is good; Residuals vs Leverage plot: there is no outliers that exceed 3 standard deviations so the model is good. Table 4 shows estimates, confidence interval and p-value of the model. Because only p-value of FGM is smaller than 0.005, every one unit increase difference in FGM is associated with a 1.84 increase in Plus Minus of James Harden. Moreover, FG3A, FGA, FG3M have no strong correlations with Plus Minus.

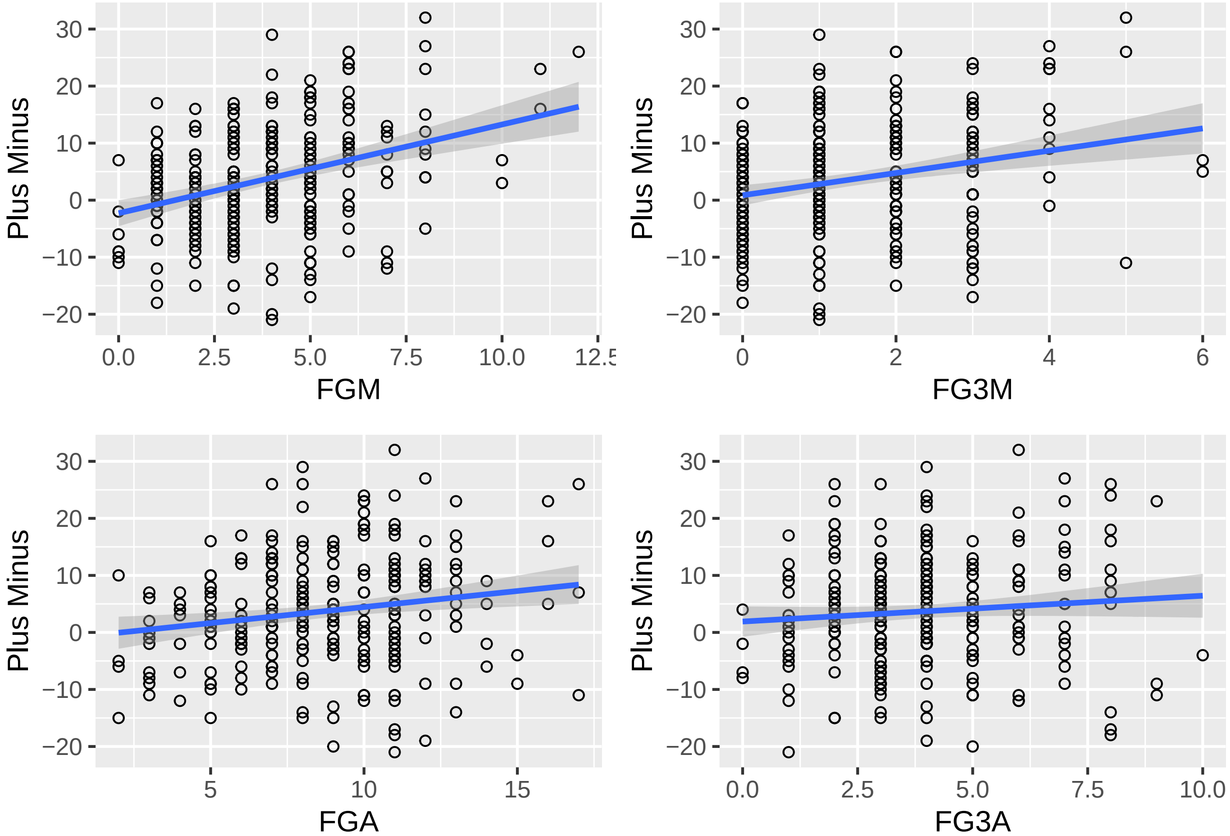

Figure 4: James Harden: Linear relationship between FGM,FGA,FG3M,FG3A and Plus Minus in Oklahoma city

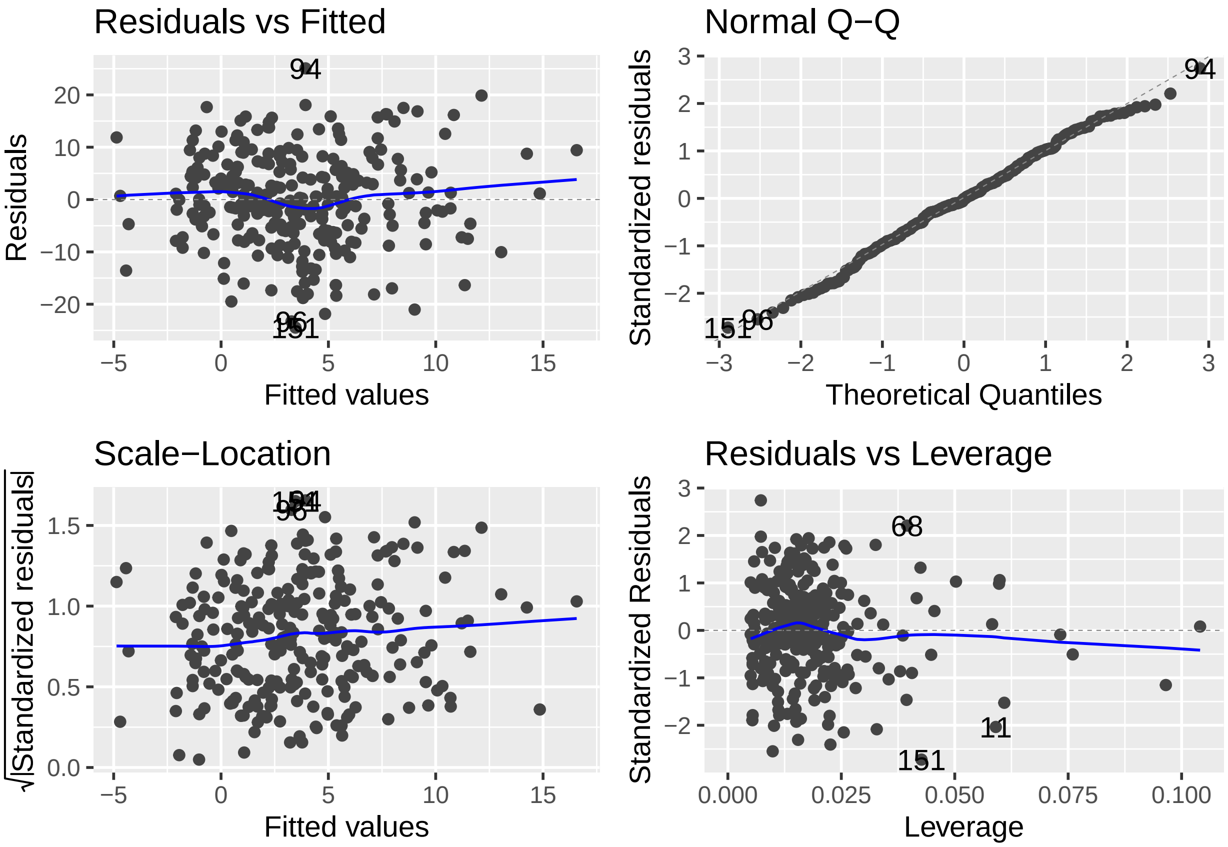

Figure 5: Model Diagnostics

Table 4: James Harden: model summary of Oklahoma

	Estimates	CI	P
(Intercept)	0.17	-3.22 ~ 3.55	0.923
FG3M	0.64	-0.79 ~ 2.07	0.379
FG3A	-0.25	-1.18 ~ 0.68	0.6
FGM	1.84	0.90 ~ 2.78	<0.001
FGA	-0.4	-1.09 ~ 0.28	0.248

4.1.2 Houston

During James Harden's time in Houston, he became an all-star player. From figure 6, we can see the linear relationship between Plus Minus and FGM, FGA, FG3M, FG3A. Field Goal Made (FGM) and 3 point Field Goal Made (FG3M) show a strong linearly increasing relationship and their slopes are higher than the others significantly. It means that FGM and FG3M could affect significantly Plus Minus of James Harden in Houston. Moreover, the number of points, that are on plot, are much more than points in Oklahoma so it means James Harden played the most number of games in Houston. Then, I built a linear model to analyze the correlations between Plus Minus and FGM, FGA, FG3M, FG3A. We can see regression model diagnostics from figure 5. Residuals vs Fitted plot: there is no pattern in the residual plot so we can assume linear relationship between the predictors and the outcome variables; QQ plot: residuals points follow the straight line so the model is good; Scale-location plot: a horizontal line with equally spread points so the model is good; Residuals vs Leverage plot: there is no outliers that exceed 3 standard deviations so the model is good. Table 5 shows estimates, confidence interval and p-value of the model. We can see p-values of FG3M, FG3A, FGM, FGA are smaller than 0.05. Therefore, every one unit increase difference in FG3M is associated with a 0.85 increase in Plus Minus of James Harden. Every one unit increase difference in FG3A is associated with a 0.60 increase in Plus Minus of James Harden. Every one unit increase difference in FGM is associated with a 1.85 increase in Plus Minus of James Harden. Every one unit increase difference in FGA is associated with a 1.13 decrease in Plus Minus of James Harden. I think when James Harden was in Houston, his shooting affected his Plus Minus significantly. James Harden was an absolute leader in Houston Rocket.

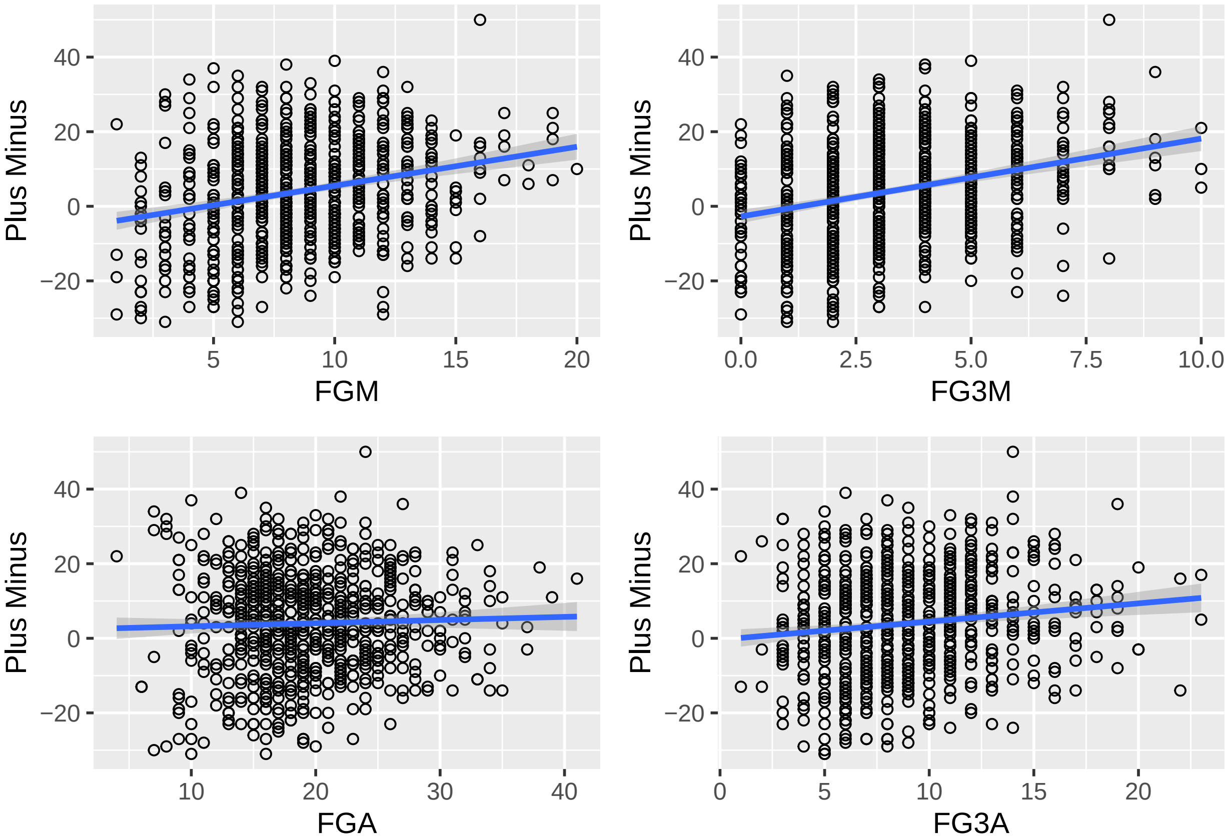

Figure 6: James Harden: Linear relationship between FGM,FGA,FG3M,FG3A and Plus Minus in Huston

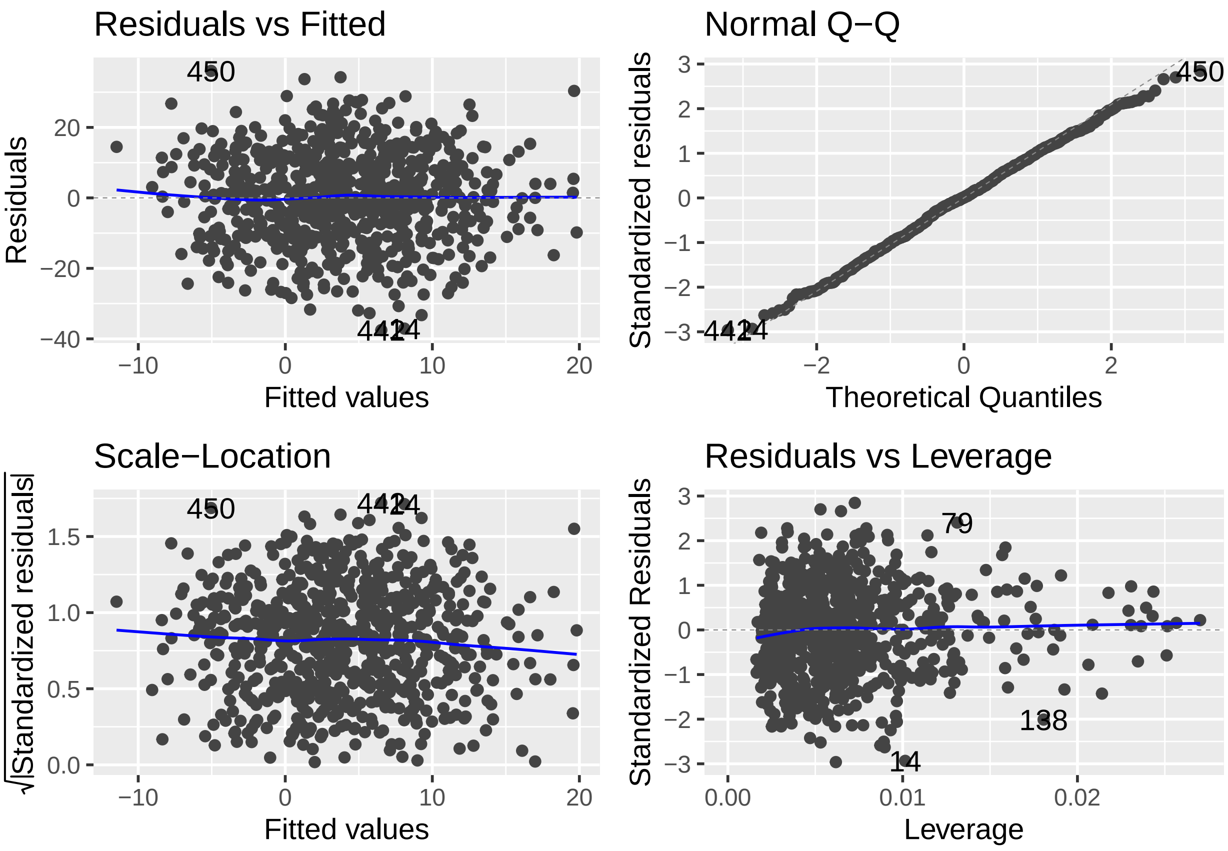

Figure 7: Model Diagnostics

Table 5: James Harden: model summary of Houston

	Estimates	CI	P
(Intercept)	2.06	-1.20 ~ 5.32	0.215
FG3M	0.85	-0.01 ~ 1.68	0.048
FG3A	0.6	0.11 ~ 1.10	0.017
FGM	1.85	1.28 ~ 2.41	<0.001
FGA	-1.13	-1.49 ~ -0.77	<0.001

4.1.3 Brooklyn

During James Harden's time in Brooklyn, he was good at setting up offenses. From figure 7, we can see the linear relationship between Plus Minus and FGM, FGA, FG3M, FG3A. Field Goal Made (FGM) and 3 point Field Goal Made (FG3M) show a strong linearly increasing relationship and their slopes are higher than the others significantly. It means that FGM and FG3M could affect significantly Plus Minus of James Harden in Brooklyn. Then, I built a linear model to analyze the correlations between Plus Minus and FGM, FGA, FG3M, FG3A. We can see regression model diagnostics from figure 8. Residuals vs Fitted plot: there is no pattern in the residual plot so we can assume linear relationship between the predictors and the outcome variables; QQ plot: residuals points follow the straight line so the model is good; Scale-location plot: a horizontal line with equally spread points so the model is good; Residuals vs Leverage plot: there is no outliers that exceed 3 standard deviations so the model is good. Table 6 shows estimates, confidence interval and p-value of the model. We can see p-values of FGM and FGA are smaller than 0.05. Therefore, every one unit increase difference in FGM is associated with a 2.28 increase in Plus Minus of James Harden and every one unit increase difference in FGA is associated with a 1.26 decrease in Plus Minus of James Harden. I think the field goal (2 point) of James Harden will affect his Plus Minus significantly when he was in Brooklyn.

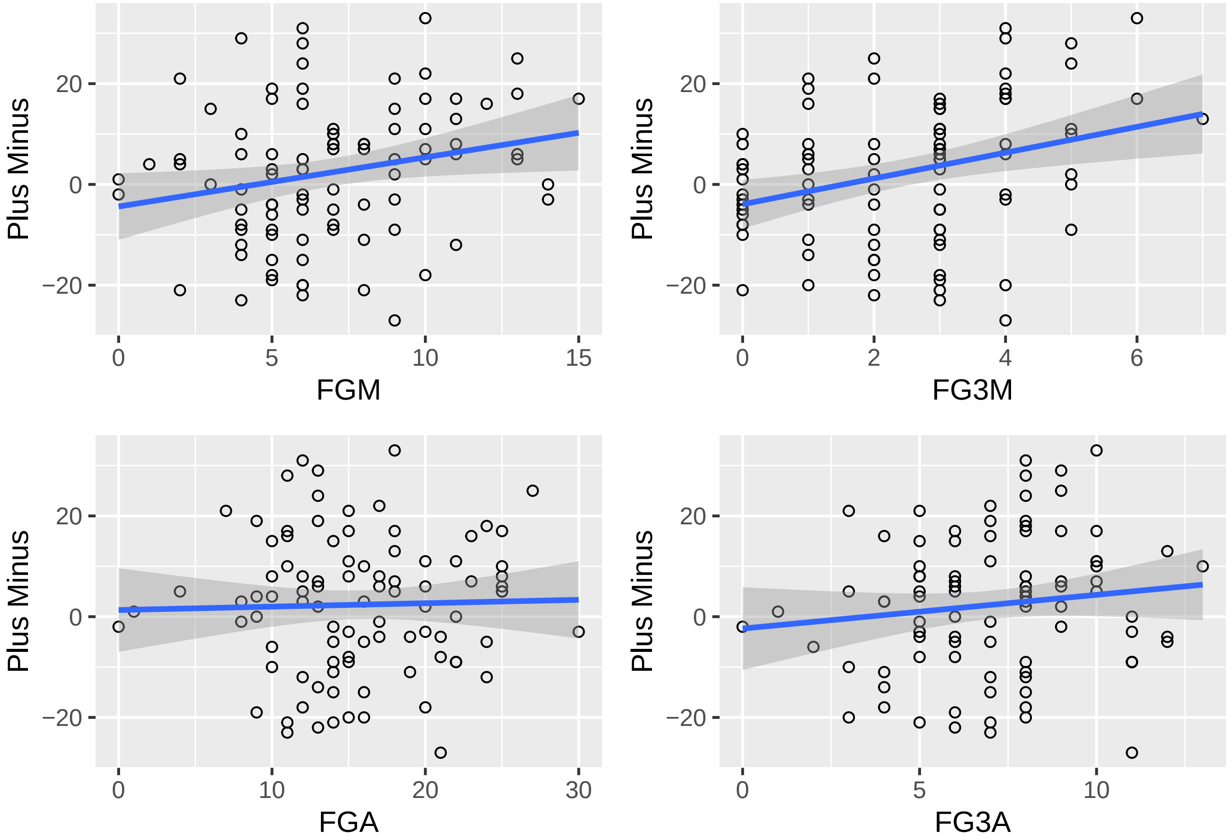

Figure 8: James Harden: Linear relationship between FGM,FGA,FG3M,FG3A and Plus Minus in Brooklyn

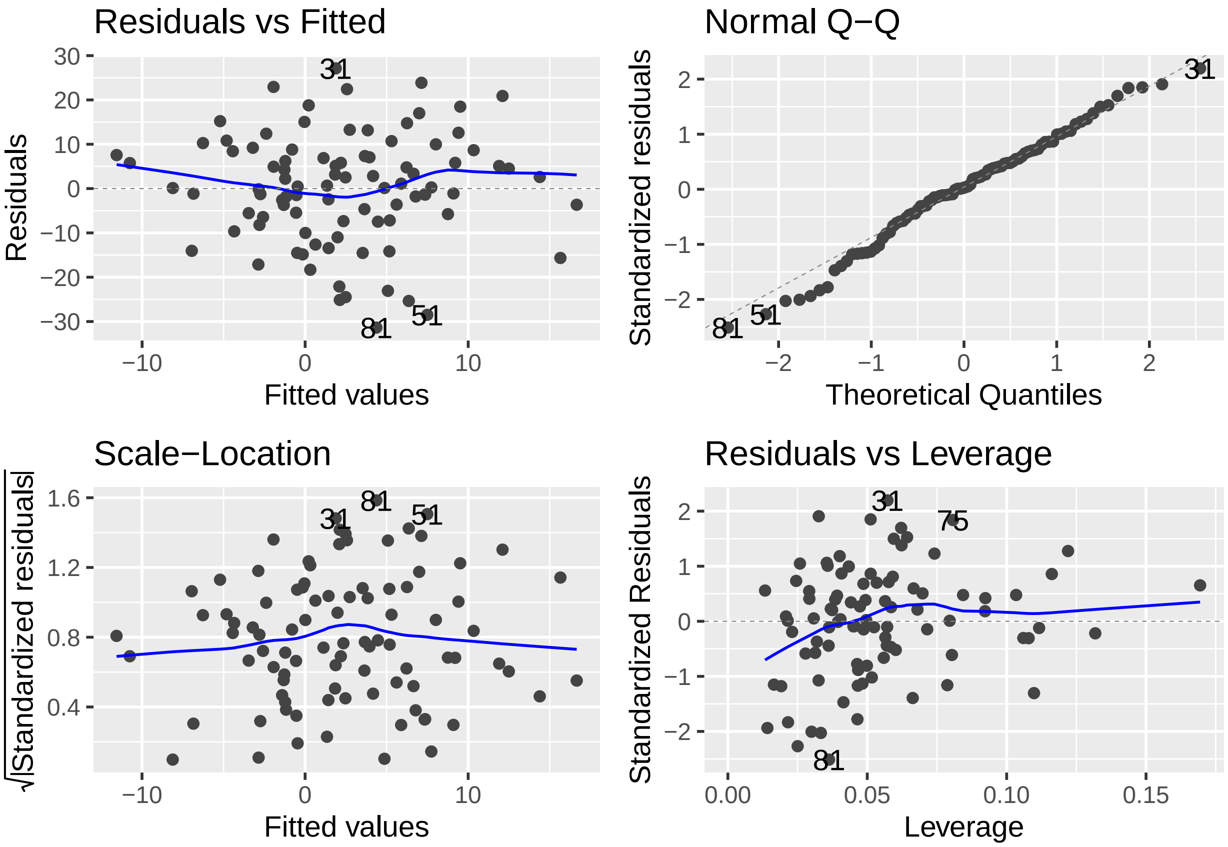

Figure 9: Model Diagnostics

Table 6: James Harden: model summary of Brooklyn

	Estimates	CI	P
(Intercept)	-0.53	-8.98 ~ 7.92	0.901
FG3M	1.12	-1.75 ~ 3.99	0.439
FG3A	0.57	-1.45 ~ 2.59	0.576
FGM	2.28	0.38 ~ 4.19	0.020
FGA	-1.26	-2.43 ~ -0.08	0.036

4.2 Kyrie Irving

4.2.1 Cleveland

During Kyrie Irving's time in Cleveland, he as a rookie showed great offensive ability. From figure 9, we can see the linear relationship between Plus Minus and FGM, FGA, FG3M, FG3A. Field Goal Made (FGM) shows a strong linearly increasing relationship and Field Goal Attempt (FGA) shows a slight linearly decreasing relationship. Then, I built a linear model to analyze the correlations between Plus Minus and FGM, FGA, FG3M, FG3A. We can see regression model diagnostics from figure 10. Residuals vs Fitted plot: there is no pattern in the residual plot so we can assume linear relationship between the predictors and the outcome variables; QQ plot: residuals points follow the straight line so the model is good; Scale-location plot: a horizontal line with equally spread points so the model is good; Residuals vs Leverage plot: there is no outliers that exceed 3 standard deviations so the model is good. Table 7 shows estimates, confidence interval and p-value of the model. We can see p-values of FGM, FG3M and FGA are smaller than 0.05. Therefore, every one unit increase difference in FG3M is associated with a 1.61 increase in Plus Minus of Kyrie Irving. Every one unit increase difference in FGM is associated with a 1.84 increase in Plus Minus of Kyrie Irving. Every one unit increase difference in FGA is associated with a 1.17 decrease in Plus Minus of Kyrie Irving. I think Kyrie Irving had a great performance in Cleveland because his shooting affects a high Plus Minus.

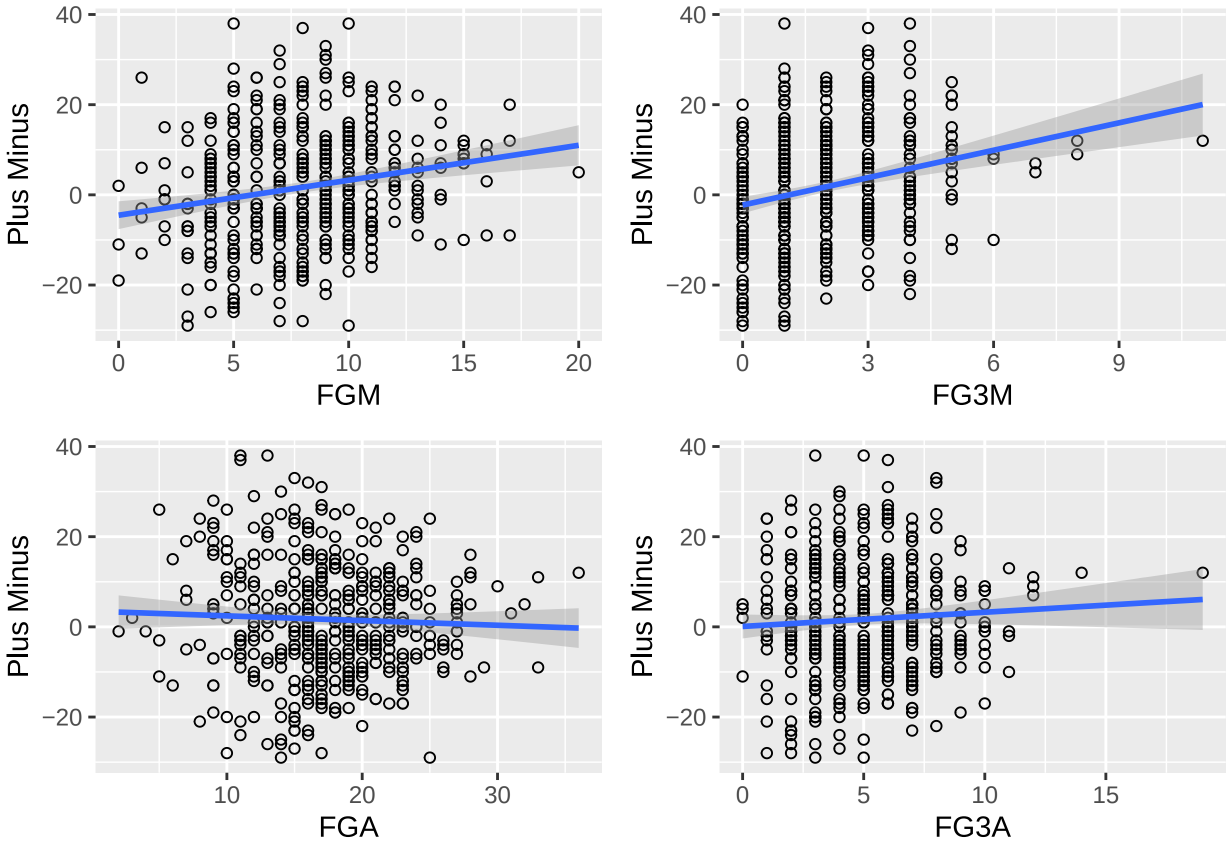

Figure 10: Kyrie Irving: Linear relationship between FGM,FGA,FG3M,FG3A and Plus Minus in Cleveland

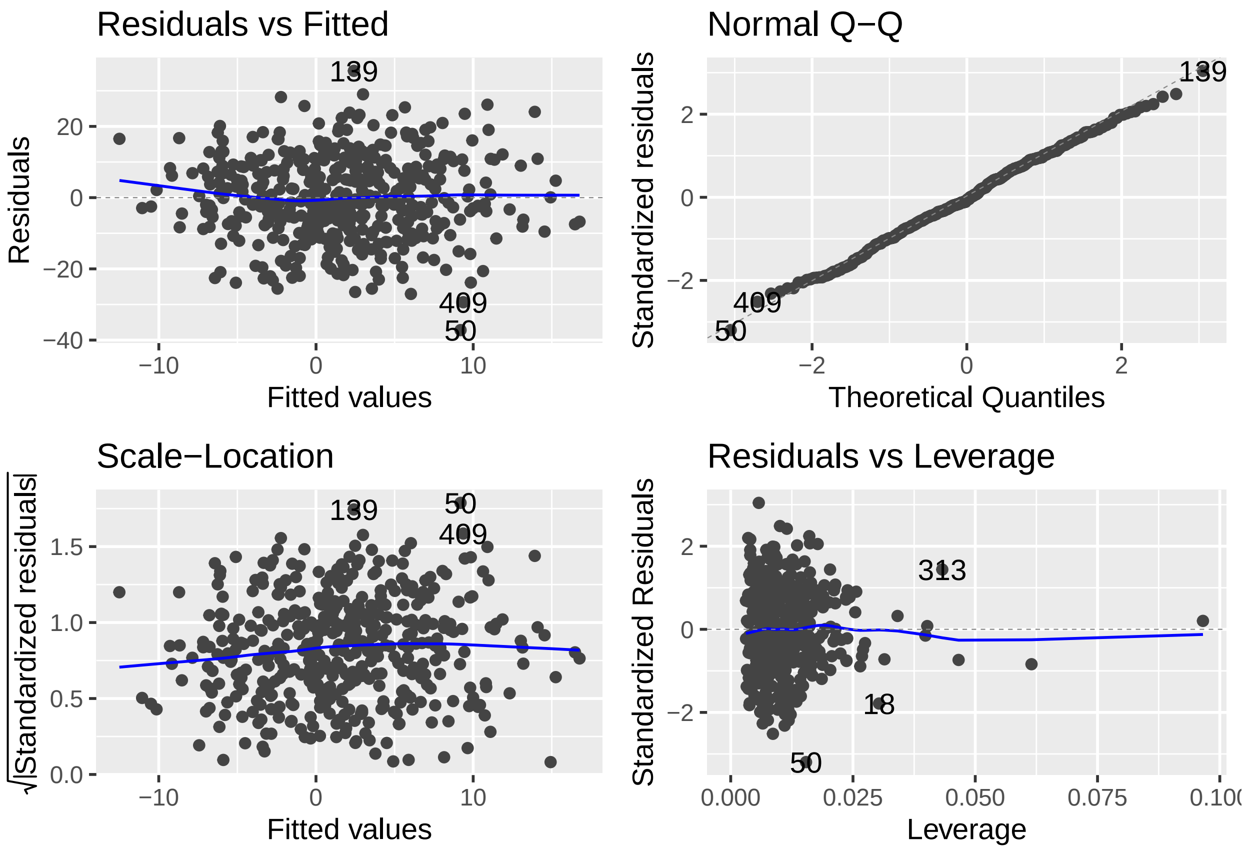

Figure 11: Model Diagnostics

Table 7: Kyrie Irving: model summary of Cleveland

	Estimates	CI	P
(Intercept)	4.84	1.02 ~ 8.67	0.013
FG3M	1.61	0.45 ~ 2.77	0.007
FG3A	-0.08	-0.86 ~ 0.69	0.834
FGM	1.84	1.23 ~ 2.46	<0.001
FGA	-1.18	-1.59 ~ -0.78	<0.001

4.2.2 Boston

During Kyrie Irving's time in Boston, he had more mature offensive ability. From figure 11, we can see the linear relationship between Plus Minus and FGM, FGA, FG3M, FG3A. Field Goal Made (FGM) and 3 Point Field Goal Made (FG3M) show a strong linearly increasing relationship. Moreover, Field Goal Attempt (FGA) shows a slight linearly decreasing relationship. Thus, FGA, FG3M and FGM could affect Kyrie's Plus Minus significantly. Then, I built a linear model to analyze the correlations between Plus Minus and FGM, FGA, FG3M, FG3A. We can see regression model diagnostics from figure 12. Residuals vs Fitted plot: there is no pattern in the residual plot so we can assume linear relationship between the predictors and the outcome variables; QQ plot: residuals points follow the straight line so the model is good; Scale-location plot: a horizontal line with equally spread points so the model is good; Residuals vs Leverage plot: there is no outliers that exceed 3 standard deviations so the model is good. Table 8 shows estimates, confidence interval and p-value of the model. We can see the p-value of FG3M equals 0.05 and the p-value of FGA are smaller than 0.05. FG3M has a slight effect on Plus Minus. Therefore, every one unit increase difference in FG3M is associated with a 2.05 increase in Plus Minus of Kyrie Irving. Every one unit increase difference in FGA is associated with a 0.83 decrease in Plus Minus of Kyrie Irving. In Boston, the Plus Minus of Kyrie Irving is 6.88 and his 3 point shooting is important for Plus Minus in the games.

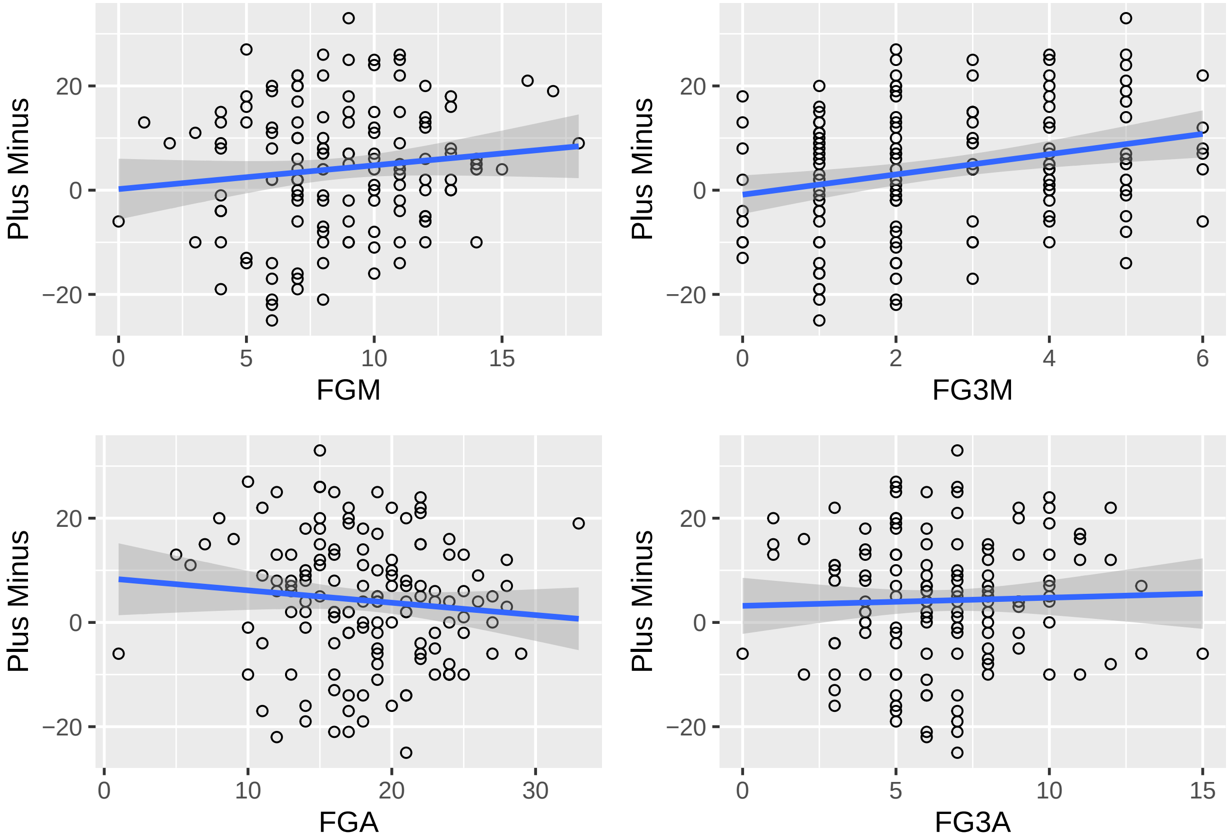

Figure 12: Kyrie Irving: Linear relationship between FGM,FGA,FG3M,FG3A and Plus Minus in Boston

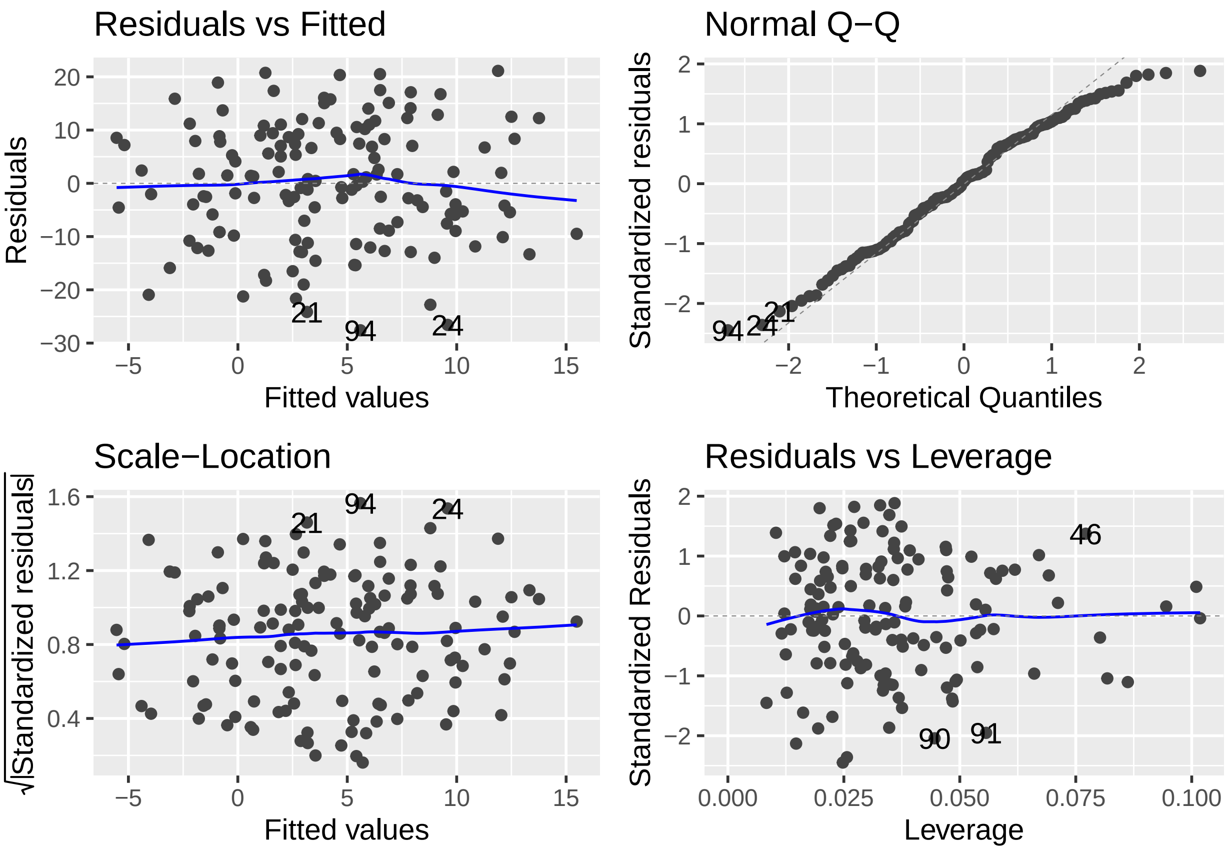

Figure 13: Model Diagnostics

Table 8: Kyrie Irving: model summary of Boston

	Estimates	CI	P
(Intercept)	6.88	-0.03 ~ 13.78	0.051
FG3M	2.05	-0.06 ~ 4.16	0.057
FG3A	-0.18	-1.57 ~ 1.22	0.803
FGM	0.94	-0.20 ~ 2.07	0.104
FGA	-0.83	-1.58 ~ -0.07	0.032

4.2.3 Brooklyn

Currently, Kyrie Irving plays an important role in the Brooklyn Net. From figure 13, we can see the linear relationship between Plus Minus and FGM, FGA, FG3M, FG3A. 3 Point Field Goal Made (FG3M) shows a strong linearly increasing relationship. Moreover, Field Goal Attempt (FGA) shows a linearly decreasing relationship. Thus, FGA, FG3M and FGM could affect Kyrie's Plus Minus significantly. Then, I built a linear model to analyze the correlations between Plus Minus and FGM, FGA, FG3M, FG3A. We can see regression model diagnostics from figure 14. Residuals vs Fitted plot: there is no pattern in the residual plot so we can assume linear relationship between the predictors and the outcome variables; QQ plot: residuals points follow the straight line so the model is good; Scale-location plot: a horizontal line with equally spread points so the model is good; Residuals vs Leverage plot: there is no outliers that exceed 3 standard deviations so the model is good. Table 9 shows estimates, confidence interval and p-value of the model. The p-values of FGM and FGA are smaller than 0.05 so every one unit increase difference in FGM is associated with a 1.57 increase in Plus Minus of Kyrie Irving and every one unit increase difference in FGA is associated with a 1.68 idecrease in Plus Minus of Kyrie Irving. In Brooklyn, Kyrie Irving's field goal is important for Plus Minus in the games.

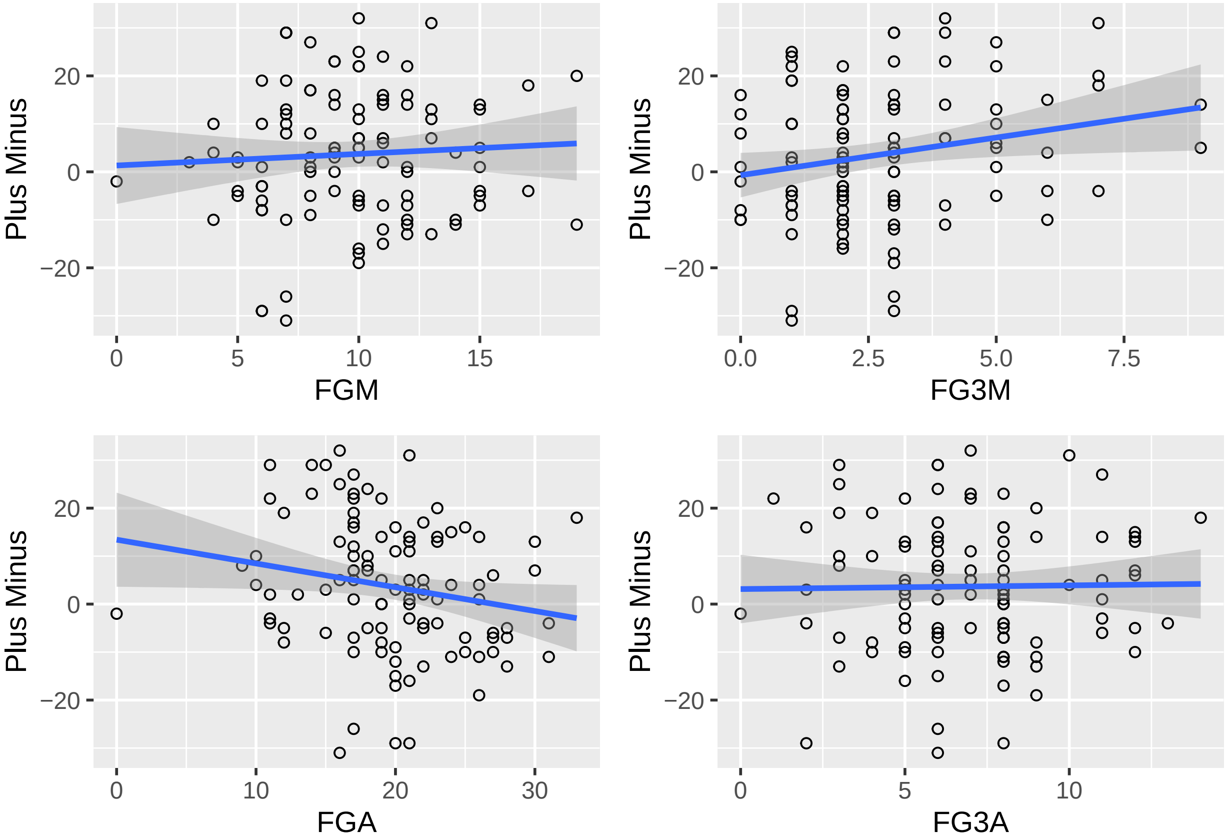

Figure 14: Kyrie Irving: Linear relationship between FGM,FGA,FG3M,FG3A and Plus Minus in Brooklyn

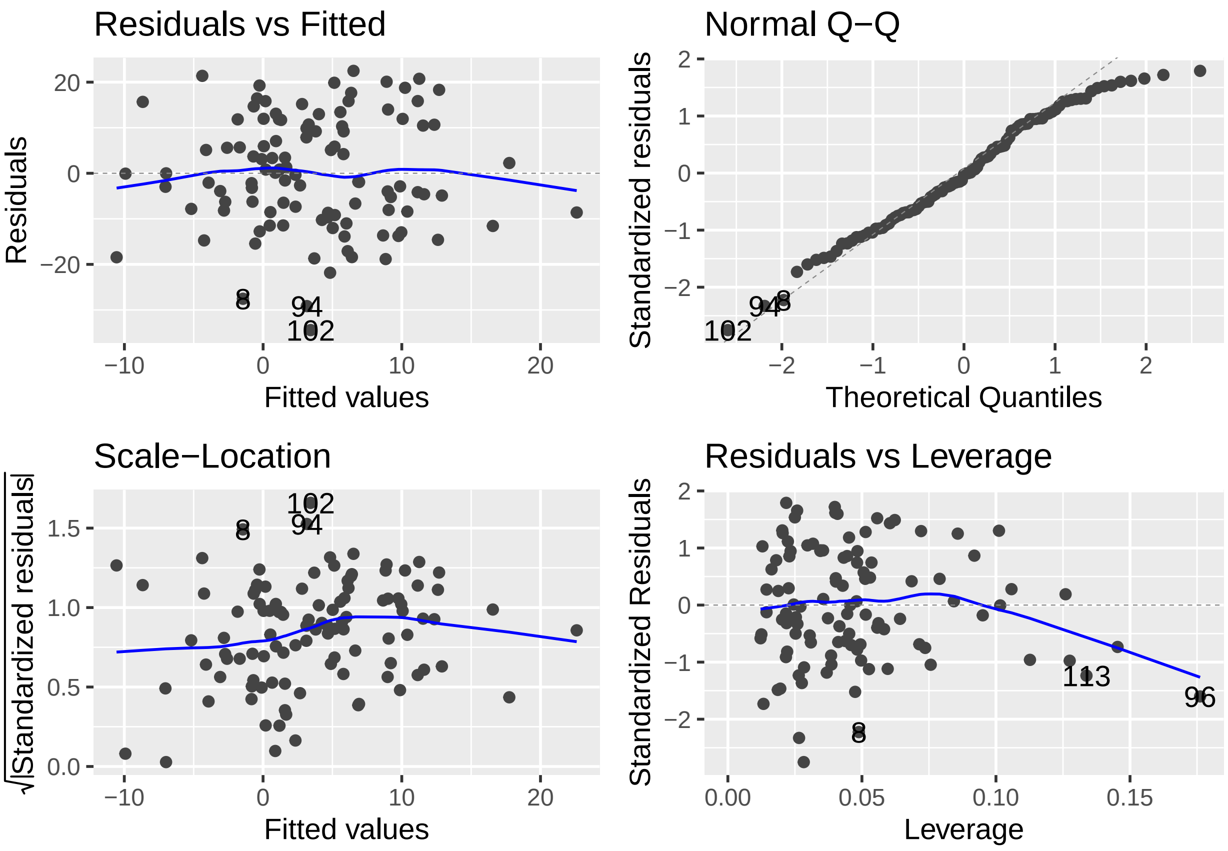

Figure 15: Model Diagnostics

Table 9: Kyrie Irving: model summary of Brooklyn

	Estimates	CI	P
(Intercept)	12.61	3.40 ~ 21.82	0.008
FG3M	0.71	-1.72 ~ 3.13	0.565
FG3A	1.00	-0.84 ~ 2.84	0.282
FGM	1.57	0.18 ~ 2.96	0.027
FGA	-1.68	-2.66 ~ -0.70	0.001

4.3 Stephen Curry

4.3.1 Golden State

Stephen Curry grew up from rookie to star in Golden State. From figure 15, we can see four plots show a strong linearly increasing relationship respectively. Therefore, FGM, FG3M, FGA and FG3A could influence Stephen Curry's Plus Minus significantly. Then, I built a linear model to analyze the correlations between Plus Minus and FGM, FGA, FG3M, FG3A. We can see regression model diagnostics from figure 16. Residuals vs Fitted plot: there is no pattern in the residual plot so we can assume linear relationship between the predictors and the outcome variables; QQ plot: residuals points follow the straight line so the model is good; Scale-location plot: a horizontal line with equally spread points so the model is good; Residuals vs Leverage plot: there is no outliers that exceed 3 standard deviations so the model is good. Table 10 shows estimates, confidence interval and p-value of the model.

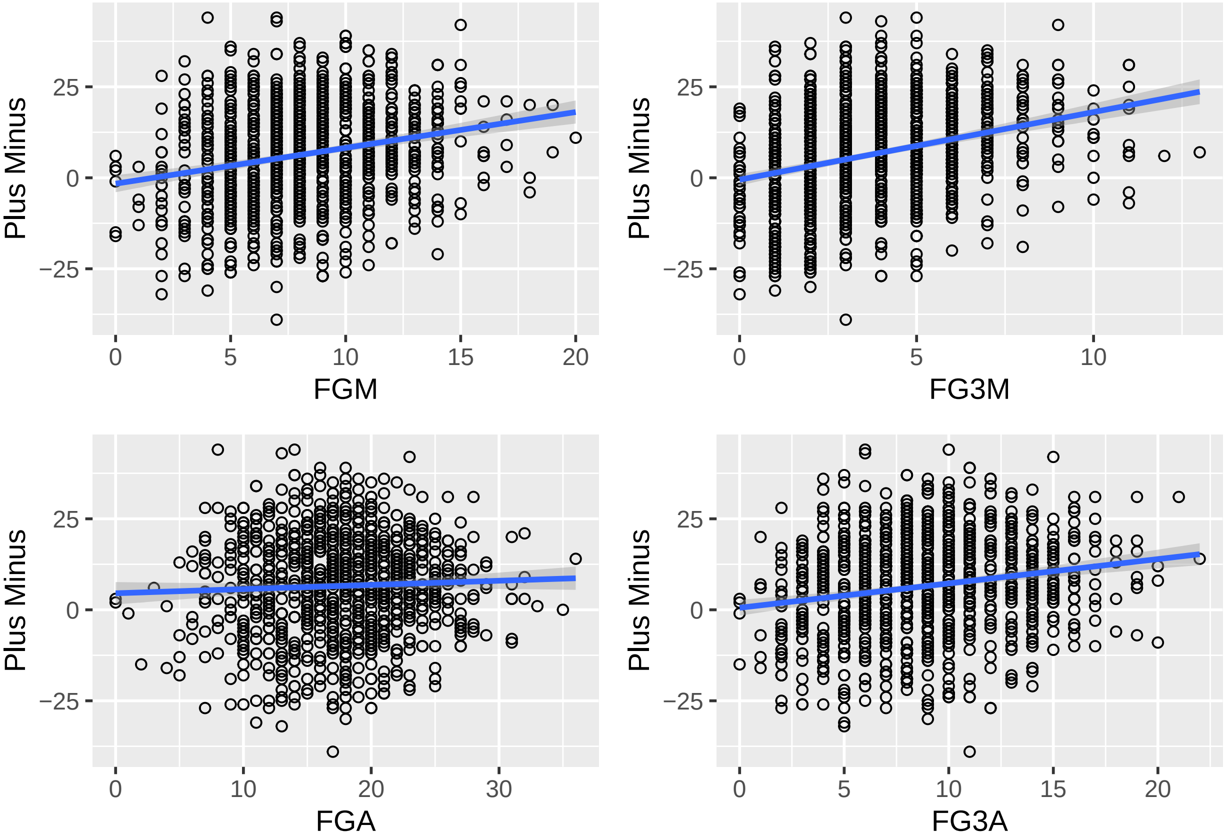

Figure 16: Stephen Curry: Linear relationship between FGM,FGA,FG3M,FG3A and Plus Minus in Golden State

Table 10: Stephen Curry: model summary of Golden State

	Estimates	CI	P
(Intercept)	6.65	3.80 ~ 9.50	<0.001
FG3M	0.58	-0.20 ~ 1.36	0.143
FG3A	1.01	-0.52 ~ 1.50	<0.001
FGM	1.88	1.30 ~ 2.45	<0.001
FGA	-1.53	-1.90 ~ -1.16	<0.001

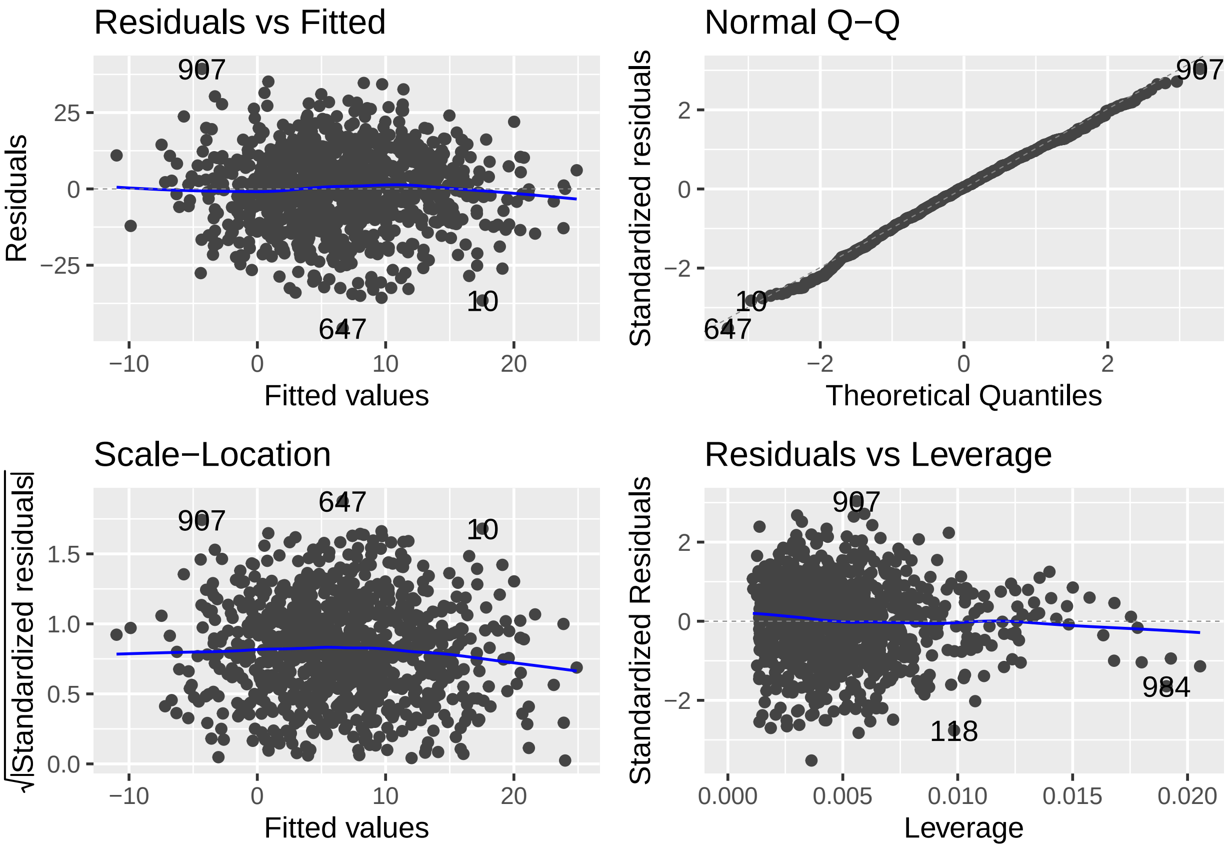

Figure 17: Model Diagnostics

4.4 Comparison

Firstly, I will compare the Plus Minus that I analyzed above. For James Harden, I think he had the best offensive ability in Houston because he got higher Plus Minus (2.06), compared with Brooklyn and Oklahoma City. In addition, when James Harden was in Houston, his FGM, FGA, FG3M, FG3A all correlate with Plus Minus. It means he is a leader in the team. For Kyrie Irving, I think he played best in Brooklyn because he got the highest Plus Minus (12.61) that doubles his Plus Minus (6.88) in Boston. Moreover, his FGA and FGM affect Plus Minus significantly. It means he always won the game by 2 point shot. For Stephen Curry, average Plus Minus (6.65) proofs that he is one of the best Point guard in league. He is not only good at 3 point shots, but also his field goals are important for games.

Secondly, I will compare other offensive data, which are 3 Point Field Goal Percentage (FG3_PCT), Field Goal Percentage (FG_PCT), Free Throw Percentage (FT_PCT), Assist (AST), Steal (STL), Turnover (TO) and Points (PTS), of James Harden, Kyrie Irving and Stephen Curry. From the first radar chart, we can see a comparison of FG3_PCT, FG_PCT and FT_PCT between them. Generally, their percentages are really closed. However, green points are cover other two points in aspect of FG3_PCT and FG_PCT. It means Stephen Curry has higher 3 Point Field Goal Percentage and Field Goal Percentage than James Harden and Kyrie Ivring. Black point is over other two points in aspect of FT_PCT so James Harden has higher Free Throw Percentage. From the second radar chart, we can see a comparison of AST, STL and TO between them. Generally, the area of James Harden is over others' area. Green point is over other two points in aspect of STL so Stephen Curry has more steals per game. Moreover, black points are over two points in aspect of AST and TO so James Harden is really good at organizing offense. Therefore, Kyrie Ivring totally has weaker shooting rate, assist, steal, turnover, compared with Stephen Curry and James harden. At last, I analyze their points in each game by a boxplot. Boxplot shows their medians of scoring. Median value of James Harden is approximate 24 points; Median value of Kyrie Irving is approximate 22 points; Median value of Stephen Curry is approximate 25 points. Upper quartile of James Harden is the highest. Therefore, I think James Harden and Stephen Curry have more average points than Kyrie Irving.

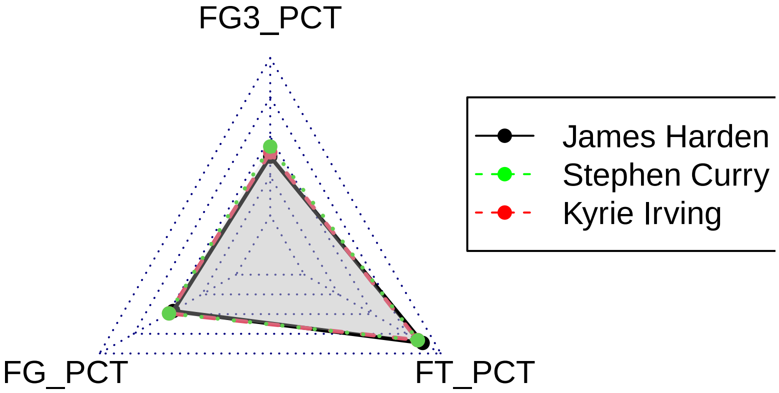

Figure 18: Offensive abilities between James Harden, Kyrie Irving and Stephen Curry

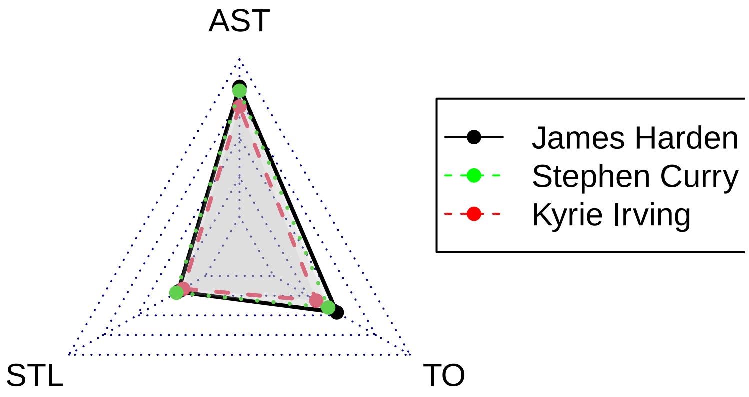

Figure 19: Offensive abilities between James Harden, Kyrie Irving and Stephen Curry

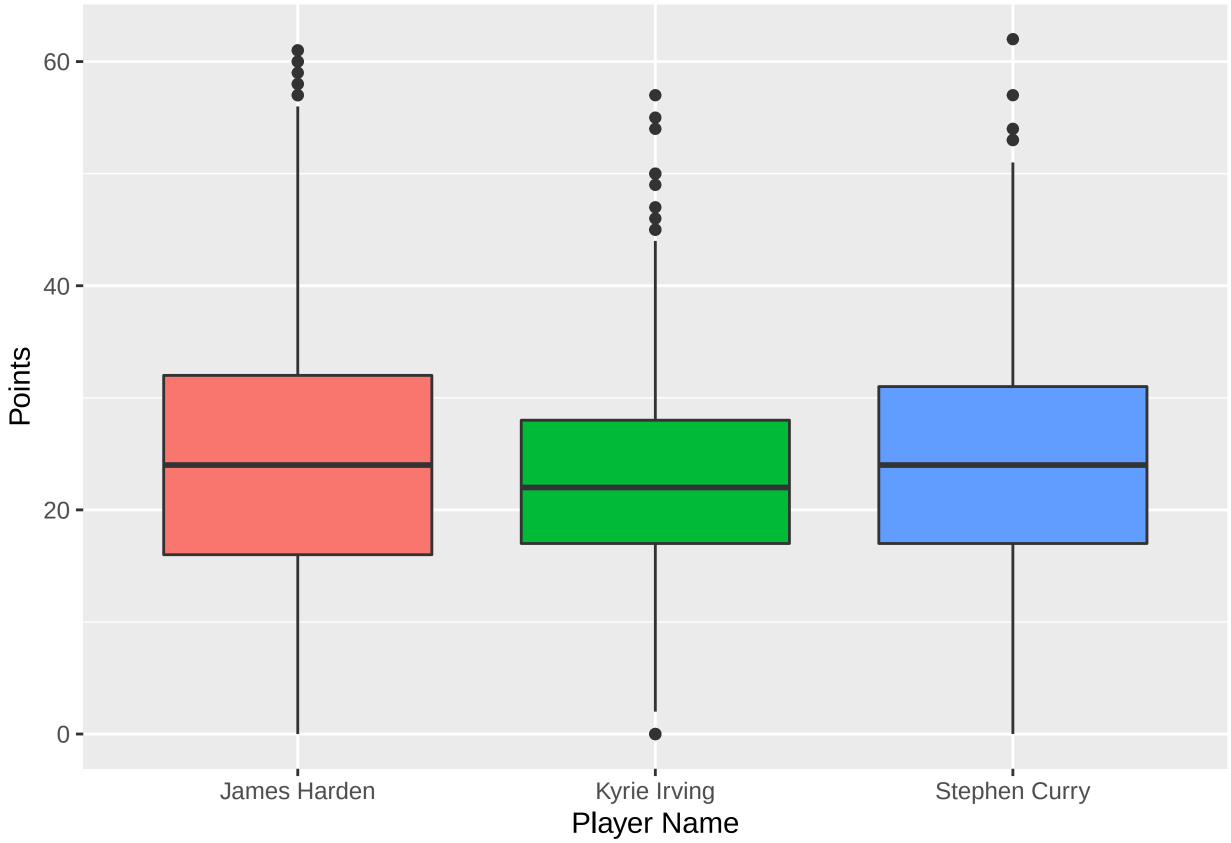

Figure 20: Scoring between James Harden, Kyrie Irving and Stepthen Curry

5. Discussion

In the result section, I analyzed all offensive abilities of James Harden, Kyrie Irving and Stephen Curry respectively. Firstly, results describe linear relationship between their Plus Minus and FGA, FGM, FG3A, FG3M. I find out James Harden became an absolute leader when he was in Houston Rocket because his FGA, FGM, FG3A, FG3M all affect his Plus Minus significantly. Kyrie Irving has the highest Plus Minus (12.61) when he was in Brooklyn and his FGA and FGM affect Plus Minus significantly. It means that he played an important role in winning the game. Stephen Curry have always been the core of Golden State Warriors. He has a great Plus Minus (6.65) and his FGA, FGM, FG3A affect Plus Minus significantly. Furthermore, I compare their 3 Point Field Goal Percentage (FG3_PCT), Field Goal Percentage (FG_PCT), Free Throw Percentage (FT_PCT), Assist (AST), Steal (STL) and Turnover (TO). I find out James Harden has prominent FT_PCT, AST and TO and Stephen Curry has prominent FG_PCT, FG3_PCT and STL. Kyrie Irving has no strengths, compared with James Harden and Stephen Curry. At last, I find out James Harden and Stephen Curry have similar median value of scoring (24 points) that is higher than Kyrie Irving's.

5.1 James Harden

From result section, when James Harden was in Houston Rocket, his FGA, FGM, FG3A, FG3M all affect his Plus Minus significantly. In fact, James Harden was named MVP in 2018 because Harden is averaging 25.3 points, 7.9 rebounds and a league-best 11.2 assists per game (Ben 2019). The 2017 - 2018 reason is his sixth season in Rocket. Therefore, I think James Harden had the best performance in Houston and he was definitely the leader of the team. His three point shot and two point shot not only influenced his own Plus Minus, but also led the team to win. However, we can see James Harden did not well in Brooklyn because of a negative Plus Minus. This is because James Harden experienced a hamstring injury in Brooklyn Net (Nick 2022). Therefore, he missed some games and reduced many offensive opportunities. Moreover, we can see James Harden has prominent free throw percentage, assist and turnover from the result section. Harden led the NBA in assists per game (11.2) in the 2016-17 season. He was good at generating others' fouls to get free throws and he needed to practice free throws a lot (Ben 2019). Therefore, he has a high free throw percentage in the games.

5.2 Kyrie Irving

From the result section, Kyrie Irving got the highest Plus Minus (12.61) in Brooklyn. It means he had the best performance in Brooklyn. Kyrie Irving's return brings the Nets some stability and they expect the championship (Tania 2022). At the beginning of the 2021 - 2022 season, Irving did not attend games because of COVID-19 vaccines. However, his return made the Brooklyn Nets go to the playoffs directly because he was on fire. Furthermore, Kyrie Irving scored 60 points, which is his career-high point in the game (Brooklyn Nets vs Orlando Magic). In the game, he shot 20 for 31 including 8 on three-pointers (Tdowd 2022). Kyrie Irving had more space to express his offensive abilities in Brooklyn. This is because Kevin Durant is a good teammate with Kyrie Irving. Durant said that I think you could see that we're both in a nice little groove right now and we want to continue (Tdowd 2022).

5.3 Stephen Curry

From the result section, Stephen Curry has always been the core of Golden State Warriors because he has a great Plus Minus. Stephen Curry won his first MVP award and led the team to the championship in the 2014 - 2015 season. Then, he led the Warriors to the NBA Finals in 2016, 2017, 2018, 2019 (C. 2021). I think Stephen Curry had strong offensive abilities and led the Golden State Warriors to become one of the strongest teams in the league. In addition, Stephen Curry has a prominent 3 Point Field Goal Percentage (FG3_PCT) and Field Goal Percentage (FG_PCT). During the 2012 - 2013 season, Stephen Curry got the NBA record for three-pointers made (272) in a regular season, and then he renewed the record in 2015 with 286, and again in 2016 with 402 (C. 2021). I think 3-pointers is a brilliant label of Stephen Curry. Stephen Curry became the all-time leader in 3-pointers made with 2799 (Afahey 2022). Currently, he is the legend of 3-pointers. In the 2022 NBA ALL-STAR game, Stephen Curry made 16 3-pointers and got 50 points (Afahey 2022).

5.4 Limitations and Future

There are 3 main limitations in the paper. Firstly, the dataset just collected information from 2004 to 2020. It misses the 2021-2022 season. I think it makes my analysis not completed because players' latest stats are important for analysis. Secondly, basketball is a game of five players. James Harden, Kyrie Irving and Stephen Curry are excellent guards but their Plus Minus are related to their teammates significantly. For example, James Harden is good at assist but his teammates always miss his passes. It can affect the Plus Minus of James Harden. There are more reasons which will affect my analysis. These reasons make my analysis inaccurate because of causation. Therefore, I think our analysis does not think about related factors and it is not comprehensive. Thirdly, treatment group and control group are needed in a compared analysis. However, I cannot let James Harden, Kyrie Irving and Stephen Curry have the same height, weight, diet and so on. Therefore, I think my analysis is inaccurate.

In the future, I think James Harden, Kyrie Irving and Stephen Curry will not only influence the whole league, but also influence a generation. They will define a new guard style in the league, guard should know how to understand defense, organize offense, field goals, 3 point field goals and so on. More importantly, the 3-pointer is their strongest weapon so everyone practices 3-point shots in the league, even a center. Moreover, they will become the deepest memories of us. For example, people will remember Stephen Curry and Golden State Warriors because they reached the NBA Finals 4 times continuously. It's just like Michael Jordan and the Chicago Bulls. James Harden, Kyrie Irving and Stephen Curry bring too much excited moment for fans, I think they will enter the hall of fame in the future.

Apendix

A Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset aims to collect, analyze and disseminate offensive abilities of point guard.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by NBA stat apartment.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The creation was funded by the NBA Official.
4. *Any other comments?*
    - TBD

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The instances represent the offensive level of piont guards. The types are: 3-point filed goal, field goal, assist steal, scoring, turnover and free throw.

2. *How many instances are there in total (of each type, if appropriate)?*

- There are 32 instances

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- The 1998 NDHS sample is a sub-sample of the new master sample of the Integrated Survey of Households (ISH) of the NSO. In this data, every region; age group and sexes was included. We could say the dataset does contain all possible instances.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

- In the raw data, the instance consists of 32 variables.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

- None.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- There is no missing individual instances.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- There are no relationships between individual instances.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- There are no recommended data splits.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- There are no errors, sources of noise, or redundancies in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The dataset is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

- There is no confidential data, and the dataset is publicly available.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

- Columns that might cause anxiety include: the education level in different regions are different, this may lead to some inferiority complex.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- The dataset entirely comprises differnet age groups and differnt sex (men and women) and different regions.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

- It is not possible to identify individuals in any way.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

- Sensitive columns may include but are not limited to: different education levels for men and women.

16. *Any other comments?*

- None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- 2004-2022 NBA games details from NBA stats website. The data is collected and provided by NBA games data (NBA Official). In this dataset.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- Software programs and manual human curation.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- Households were systematically sampled in urban areas to distribute the NDHS sample throughout the sampling area, while compact clustering was used in rural areas to facilitate field operations.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

- Workers in NBA and they can gain monthly salary around 10k dollars.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- The data was collected in 2020.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- Ethical review processes were not conducted.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- We obtained the data via the NBA official website.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- No.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

- No.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

- The mechanism to revoke their consent was not provided.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- The analysis of the potential impact of the dataset and its use on data subjects has not conducted.

12. *Any other comments?*

- None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

- The data was appearing as a table originally in the PDF report. We obtain the data information in this PDF table and convert it into the data frame in R by pdftools package for R.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

- The raw data obtained from the PDF is saved in inputs/data/just_page_i.csv and inputs/data/just_page_i1.csv.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

- R Software is avalaible at https://www.R-project.org/

4. *Any other comments?*

- None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

- The dataset has not been used for other tasks yet.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

- TBD

3. *What (other) tasks could the dataset be used for?*

- The dataset can be used for analyze the different education level status.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- The process of cleaning data is specific to only this table in the original PDF report. This is not suitable in other tables.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- The dataset is not suitable for any other purposes except the education level in the aspects of age group, sex and regions.

6. *Any other comments?*

- None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

- No, this dataset if openly available.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- The dataset will be distributed using Github.

3. *When will the dataset be distributed?*

- The dataset will be distributed in April 2022.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- The dataset will be released under the MIT license.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- There are no restrictions.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

- None.

7. *Any other comments?*

- None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

- Yingxuan Shi

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- This can be contacted by Github.

3. *Is there an erratum? If so, please provide a link or other access point.*

- There is no erratum available.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- No, the dataset will not be updated.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

- This dataset was collected by the quastionnaires which were filled by people in Phillipine in 1998. There are no applicable limits.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

- The older versions of the dataset are not hosted. The dataset somsumers could be able to check the dataset by github.

Afahey. 2022. "Stephen Curry Makes 16 3-Pointers, Scores 50 to Earn All-Star MVP Honors." *NBA.com/Warriors*. https://www.nba.com/warriors/news-blogs/curry-wiggins-2022-all-star-game-20220220.

Ben. 2019. "30 Games for 30: James Harden's Best Performances in Houston." *Gannett Satellite Information Network*. https://rocketswire.usatoday.com/2019/08/26/30-games-for-30-james-hardens-best-performances-in-houston/.

C., Michael. 2021. "Stephen Curry Overtakes Ray Allen for NBA's All-Time 3-Point Lead." *NBA.com*. https://www.nba.com/news/stephen-curry-tracker-all-time-3s-record.

Chris. 2020. "What Are Basketball Positions?" *PRO TIPS by DICK'S Sporting Goods*. https://protips.dickssportinggoods.com/sports-and-activities/basketball/court-essentials-basketball-positions-explained.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Global, NBA. 2021. "About the NBA." https://global.nba.com/aboutnba/.

Nakazawa, Minato. 2022. *Fmsb: Functions for Medical Statistics Book with Some Demographic Data*. https://CRAN.R-project.org/package=fmsb.

Nick. 2022. "Brooklyn Nets' James Harden Misses Third Straight Game Because of Hamstring Injury Amid Trade Speculation." *ESPN Internet Ventures*. https://www.espn.com/nba/story/_/id/33244497/brooklyn-nets-james-harden-miss-third-straight-game-hamstring-injury-amid-trade-speculation.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Ren, Kun, and Kenton Russell. 2021. *Formattable: Create 'Formattable' Data Structures*. https://CRAN.R-project.org/package=formattable.

Tania. 2022. "Kyrie Irving Makes His Brooklyn Return." *The New York Times*. https://www.nytimes.com/2022/03/28/sports/basketball/kyrie-irving-brooklyn-nets-return.html.

Tdowd. 2022. "Brooklyn Nets' Kyrie Irving Scores Franchise-Record 60 Points." *Brooklyn Nets*. https://www.nba.com/nets/news/feature/2022/03/15/brooklyn-nets-kyrie-irving-scores-franchise-record-60-points.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2021. *Devtools: Tools to Make Developing r Packages Easier*.

Zhu, Hao. 2022. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.