# Data Exploration Project

32602243
Shu-An Lin

1. Introduction

As a basketball fan, I have been watching NBA games for over 20 years. There are over 400 players in the league, and in every season, some new stars come in while some players could be eliminated; however, some superstars not only still survive, but also always rank among the best in various statistics in such a competitive league, LeBron James is one of such a player. He is 37 years old now, but it seems that his physical age does not bother him; hence, I am interested with verifying if his stats are not impact by aging or other factors.

While there is lot of data analysis which provides people to access LeBron James 's stats season by season; there are few comprehensive comparisons of his career stats, including how early-career performance compares to now, etc. From the perspective of an NBA audience, I could expect how LeBron James may change his style of playing ball, and even whether his team will change his tactics and adjust any play strategies for him in the future.

So, I would like to ask three questions to understand if there is any change in his career integration data, and to obtain an at-a-glance graph through data visualization to answer these questions-

I. What is the trend of Lebron James's yearly data including total goals, rebounds and blocks?
II. Has there been any change in shooting distance and shooting percentage in the first five years (2003-2008) and the recent five years (2016-2021) of LeBron James's career?
III. How do different the teams affect LeBron James's performance?

2. Data Wrangling

In order to answer the questions above, there are two datasets would be needed, the steps in data wrangling, including how to access and choose the data from the sources, data cleaning and data transformations will be display as follows. All the steps are with R or Excel.
Since there are two different datasets, the processes will be divided to two parts- A and B to illustrate.

A. LeBron James's total stats in regular session and playoff games from 2003 to 2021
Link [Appx. A]: The 2020-2021 season data: https://www.basketball-reference.com/players/j/jamesle01/gamelog/2021

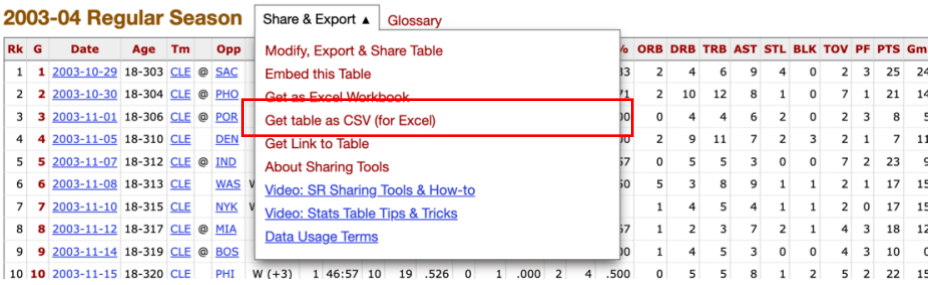
a. Data Description
Take 2020-2021 season's data as example. There are 73 games in regular session in 2020-2021 session, and 6 games in playoff session. In each game, it recorded field goals, rebounds, blocks, steals, etc in the stats. To sum up, in a season, there are approximately 80 to 100 games which depend on the

standings. For overall stats from the year 2003 to 2021, the dataset shape is 1705 rows x 30 columns.

b. Steps

Step 1: Obtain data and save into .csv files

From the link above, there are every year's records separated in different pages, as the graph shown below (Picture 2.1), choosing "Get table as CSV (for Excel)", then the format would turn as csv format (Picture 2.2), next simply copy and paste the data to a blank Numbers file (Mac OS application), and export it as .csv format file.




Picture 2.1

```
Rk,G,Date,Age,Tm,,Opp,,GS,MP,FG,FGA,FG%,3P,3PA,3P%,FT,FTA,FT%,ORB,DRB,TRB,AST,STL,BLK,TOV,PF,PTS,GmSc,+/-
1,1,2003-10-29,18-303,CLE,@,SAC,L (-14),1,42:50,12,20,.600,0,2,.000,1,3,.333,2,4,6,9,4,0,2,3,25,24.7,-9
2,2,2003-10-30,18-304,CLE,@,PHO,L (-9),1,40:21,8,17,.471,1,5,.200,4,7,.571,2,10,12,8,1,0,7,1,21,14.7,-3
3,3,2003-11-01,18-306,CLE,@,POR,L (-19),1,39:10,3,12,.250,0,1,.000,2,2,1.000,0,4,4,6,2,0,2,3,8,5.0,-21
4,4,2003-11-05,18-310,CLE,,DEN,L (-4),1,41:06,3,11,.273,0,2,.000,1,1,1.000,2,9,11,7,2,3,2,1,7,11.2,-3
5,5,2003-11-07,18-312,CLE,@,IND,L (-1),1,43:44,8,18,.444,1,2,.500,6,7,.857,0,5,5,3,0,0,7,2,23,9.0,-7
6,6,2003-11-08,18-313,CLE,,WAS,W (+13),1,44:30,8,19,.421,0,0,,1,4,.250,5,3,8,9,1,1,2,1,17,15.7,+14
7,7,2003-11-10,18-315,CLE,,NYK,W (+14),1,33:39,7,12,.583,3,3,1.000,0,0,,1,4,5,4,1,1,2,0,17,15.8,+20
8,8,2003-11-12,18-317,CLE,@,MIA,L (-5),1,42:40,6,15,.400,2,5,.400,4,6,.667,1,2,3,7,2,1,4,3,18,12.8,-4
9,9,2003-11-14,18-319,CLE,@,BOS,L (-9),1,35:36,3,12,.250,2,4,.500,2,4,.500,1,4,5,3,0,0,4,3,10,0.8,-1
10,10,2003-11-15,18-320,CLE,,PHI,W (+3),1,46:57,10,19,.526,0,1,.000,2,4,.500,0,5,5,8,1,2,5,2,22,15.6,+8
11,11,2003-11-18,18-323,CLE,,LAC,W (+8),1,39:25,6,16,.375,0,2,.000,2,4,.500,2,5,7,8,0,1,3,2,14,9.8,+1
12,12,2003-11-19,18-324,CLE,@,WAS,L (-11),1,39:09,10,18,.556,2,4,.500,6,8,.750,2,5,7,8,2,1,4,2,28,25.0,-5
13,13,2003-11-21,18-326,CLE,,MIN,L (-14),1,35:18,8,19,.421,0,2,.000,3,3,1.000,2,7,9,2,1,0,1,4,19,12.2,-5
14,14,2003-11-22,18-327,CLE,@,ATL,L (-9),1,42:18,3,16,.188,0,1,.000,9,10,.900,2,4,6,4,2,1,3,1,15,9.3,-6
15,15,2003-11-26,18-331,CLE,@,NOH,L (-10),1,38:48,5,17,.294,1,4,.250,4,4,1.000,2,6,8,9,0,0,4,1,15,10.2,-13
16,16,2003-11-28,18-333,CLE,@,DET,L (-4),1,36:27,2,8,.250,0,2,.000,2,3,.667,0,2,2,7,0,0,4,3,6,1.1,-10
17,17,2003-11-29,18-334,CLE,,MEM,L (-7),1,54:41,14,28,.500,3,6,.500,2,3,.667,2,14,16,7,1,1,8,1,33,22.4,-10
18,18,2003-12-02,18-337,CLE,@,DEN,L (-12),1,32:31,6,19,.316,2,3,.667,5,7,.714,3,3,6,5,2,0,2,4,19,12.2,-12
19,19,2003-12-03,18-338,CLE,@,LAC,L (-10),1,34:05,2,13,.154,0,3,.000,0,0,,4,2,6,8,2,0,5,2,4,0.9,+3
20,20,2003-12-06,18-341,CLE,,ATL,W (+10),1,36:04,4,12,.333,0,2,.000,0,0,,1,5,6,5,2,1,3,2,8,5.8,-5
21,21,2003-12-09,18-344,CLE,,TOR,L (-7),1,37:03,7,16,.438,0,0,,4,4,1.000,0,5,5,3,1,0,2,2,18,11.4,-2
```

Picture 2.2

Step 2: Combine files

From step1, I already have had the dataset; however, each season's data separated into different csv files. As the result, in this step, each season's data have to combine together by using R libraries, "readr" and "data.table", and "for loop" to access each file then write and save all data into a new .csv file.

Step 3: Transform into a data frame

After combing files into one, we need to transform the data from a .csv file into a data frame to have better view and easier to clean the data by using R library "tidyverse", read_csv function.

Step 4: Clean data

Next step is to clean the data, choosing those data that LeBron James played. Since LeBron James did not attend every game, there are some "Inactive", "Did Not Dress" or "Did Not Play" records in data which need to be deleted. In addition, there are some missing values (NA) existed in the data frame as Picture 2.3 shown, the missing values may exist in column "RK", "G" or the

column that displayed '@' symbol. In order to avoid deleting the games' records (blue box), first only delete these 3 columns, then to clean "Inactive", "Did Not Dress" or "Did Not Play" records (red box).

| | Rk | G | Date | Age | Tm | | Opp | | GS | MP | FG | FGA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1657 | 81 | 81 | 2018-04-09 | 33-100 | CLE | @ | NYK | W (+14) | 1 | 38:38 | 7 | 16 |
| 1658 | 81 | NA | 2019-04-07 | 34-098 | LAL | NA | UTA | W (+4) | Inactive | Inactive | Inactive | Inactive |
| 1659 | 82 | 79 | 2004-04-14 | 19-106 | CLE | @ | NYK | W (+10) | 1 | 35:05 | 8 | 17 |
| 1660 | 82 | 80 | 2005-04-20 | 20-111 | CLE | @ | TOR | W (+9) | 1 | 48:00 | 10 | 23 |
| 1661 | 82 | NA | 2006-04-19 | 21-110 | CLE | NA | ATL | W (+1) | Inactive | Inactive | Inactive | Inactive |
| 1662 | 82 | 78 | 2007-04-18 | 22-109 | CLE | NA | MIL | W (+13) | 1 | 39:58 | 9 | 18 |
| 1663 | 82 | NA | 2008-04-16 | 23-108 | CLE | NA | DET | L (-10) | Inactive | Inactive | Inactive | Inactive |
| 1664 | 82 | NA | 2009-04-15 | 24-106 | CLE | NA | PHI | L (-1) | Did Not Dress | Did Not Dress | Did Not Dress | Did Not Dres |
| 1665 | 82 | NA | 2010-04-14 | 25-105 | CLE | @ | ATL | L (-16) | Inactive | Inactive | Inactive | Inactive |
| 1666 | 82 | NA | 2011-04-13 | 26-104 | MIA | @ | TOR | W (+18) | Did Not Play | Did Not Play | Did Not Play | Did Not Play |
| 1667 | 82 | NA | 2013-04-17 | 28-108 | MIA | NA | ORL | W (+12) | Inactive | Inactive | Inactive | Inactive |
| 1668 | 82 | NA | 2014-04-16 | 29-107 | MIA | NA | PHI | L (-13) | Inactive | Inactive | Inactive | Inactive |
| 1669 | 82 | NA | 2015-04-15 | 30-106 | CLE | NA | WAS | W (+5) | Inactive | Inactive | Inactive | Inactive |
| 1670 | 82 | NA | 2016-04-13 | 31-105 | CLE | NA | DET | L (-2) | Did Not Dress | Did Not Dress | Did Not Dress | Did Not Dres |
| 1671 | 82 | NA | 2017-04-12 | 32-103 | CLE | NA | TOR | L (-15) | Inactive | Inactive | Inactive | Inactive |

Picture 2.3

After cleaning, the dataset shape turns to be 1518 rows X 27 columns.

That are all the data wrangling works for part A.

B. LeBron James's shooting distance in regular session and playoff games from 2003 to 2021
Link [Appx. B]:
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2020-21&PlayerID=2544&ContextMeasure=FGA&Season=2020-21&section=player&sct=plot

a. Data Description
Take 2020-2021 data as example, in the dataset, it records every attempt that LeBron James made including the exact location in x-y coordination system, shot zone, shot range, shot distance and shot type, etc.
In a season, there are approximately 1500 shooting attempts which also depend on the conditions and states of LeBron James in each season. In total, the dataset's shape from 2003 to 2021 is 25602 rows x 24 columns.

b. Steps
Step 1: Obtain data and save as .csv files
From the link above, we can access each season's data; however, the data displayed on the webpage is not the one we need. The complete data is hidden at the backend of webpage. By using the method from article "How to download NBA shot data with R" [1], it allows us able to obtain the whole shooting data, then save it into multiple files separated by seasons.

Step 2 to step 3 are as the same as part A, combining the files and transforming whole data into a data frame by R libraries, "readr", "data.table" and "tidyverse".

It is because the data was only recorded the stats that LeBron James had played, there is no as the record of “Inactive”, “Did Not Dress” and “Did Not Play” as part A; however, the data would be checked in the next section.

Step 4: Add columns
To answer question 2- “Has there been any change in shooting distance and shooting percentage in the first five years (2003-2008) and the recent five years (2016-2021) of LeBron James’s career?”, the first five years’ [Appx. c]data and the recent five years’ [Appx. C] data have to mark by adding columns with excel, marking 2003-2008 season data as “First_five_year”, 2016-2021 season as “Recent_five_year”, as for other mark as “n/a”.

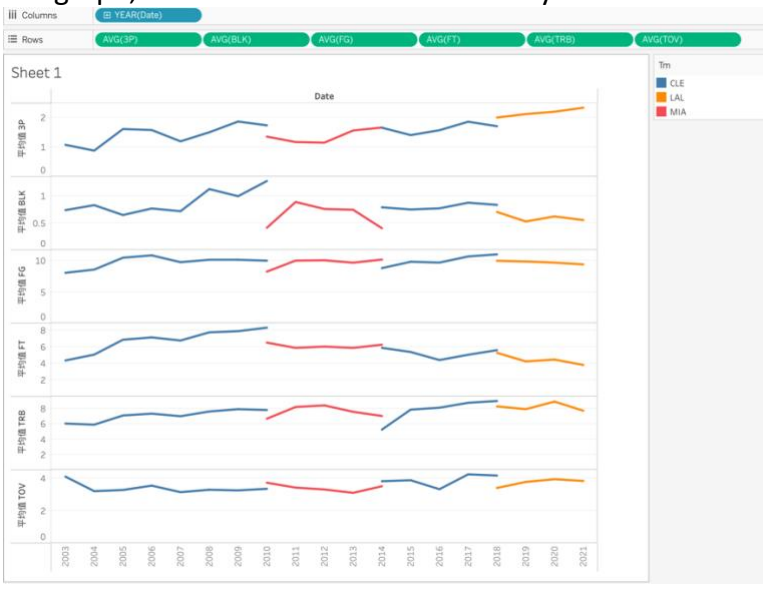
That are all the data wrangling works for part B.

## 3. Data Checking

After data wrangling section, the both datasets should be clean to use; however, these data still need to be checked if there are some missing values, or incorrect values existed. In this section, the tool I use is Tableau, by visualising the data initially to have quick over view of the dataset.
Since there are two different contents of datasets, in this section I still separate as two part to check these two datasets.

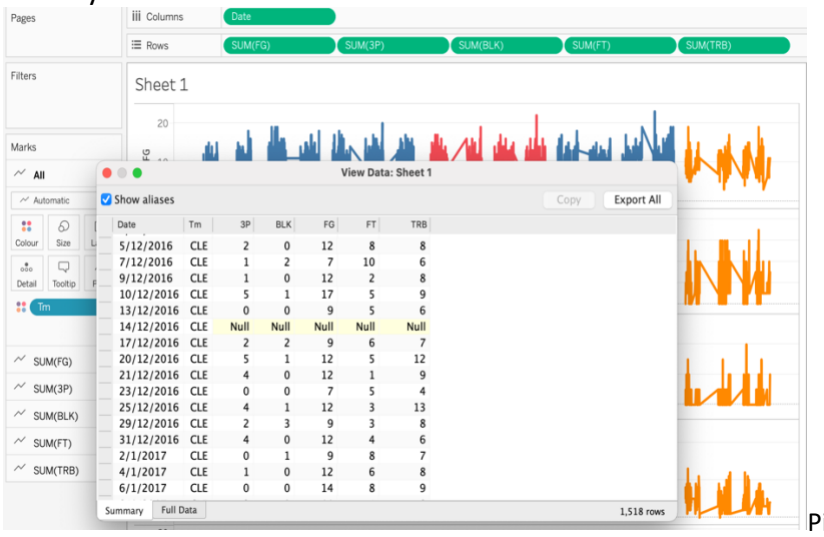
A. LeBron James’s total stats in regular session and playoff games from 2003 to 2021.

- Check teams matched to the season that LeBron James played for
  Check against the data I'm interested in, 3P, BLK, FG, FT, TRB, TOV, etc [appx. D].
  As Picture 3.1, take year data as columns, and these average stats as rows, the graph, the teams are matched to the years.

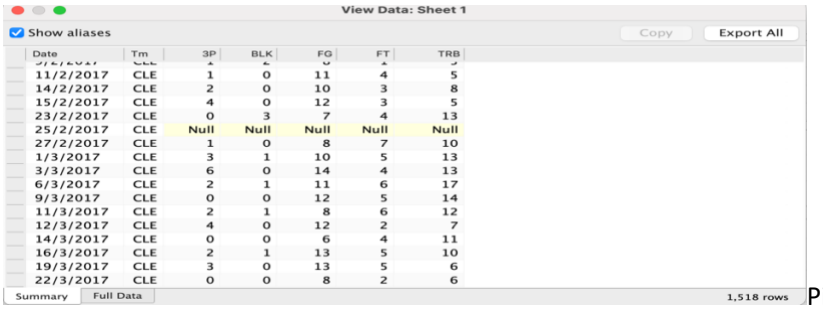


Picture 3.1

- Check if there are missing values in the dataset
  Although in the data wrangling section, I deleted the records of “Inactive”, “Did Not Dress” and “Did Not Play”, there still two null values existed as shown as Picture 3.2 and Picture 3.3 below. By deleting these two days’ stats directly in csv file to clean the data.



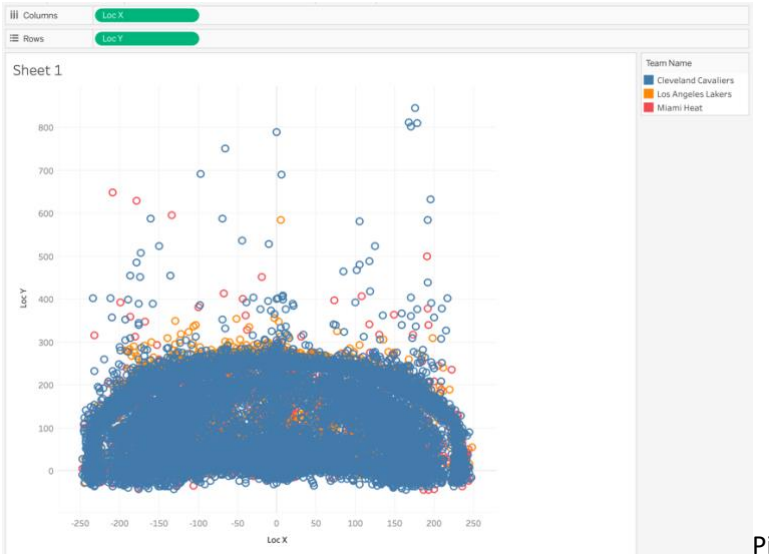
Picture 3.2



Picture 3.3

B. LeBron James’s shooting distance in regular session and playoff games from 2003 to 2021
   - Check outlier
     Sometimes basketball player may attempt shoot from backcourt when the time is going to run out; however, this kind of attempts are almost “lucky ball”, and cannot present the real shooting ability; therefore, the kind of data should be deleted, the tool used by Excel. [appx. E]

   - Change teams’ names
     Picture 3.4 shows that the shooting attempts in different locations; however, the data in blue colour is much than others. It is because that LeBron James played for Cleveland Cavaliers twice and he played for this team also longer than other two teams. One was from 2003 to 2010, and other was from 2014 to 2018. Based on the question I would like to figure out, the teams and ages are the important factors; therefore, the data in the Cavaliers stats for different periods should be distinguished. Except for the problem above, there is no discontinuous or missing value in stats.



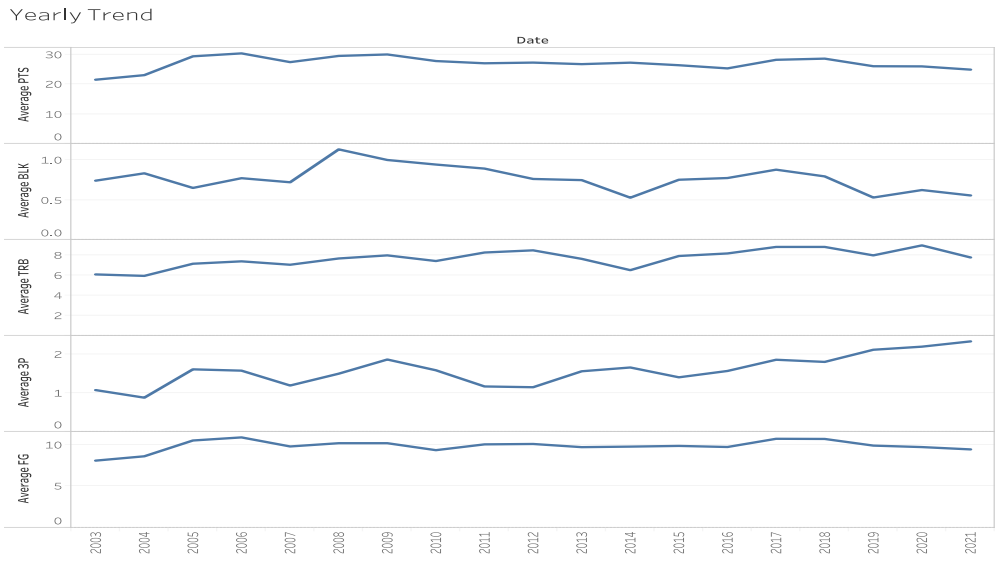
Picture 3.4

For dealing with it, changing these two periods 2003-2010 and 2014 to 2018 in Cavaliers as 03-10_Cavaliers and 14-18_Cavealiers by re-naming these columns in csv file.

Since the same issue existed in part A, CLE changed to 03-10_CLE and 14-18_CLE by using the same method above to rename them.
Now, the dataset should be ready to explore.

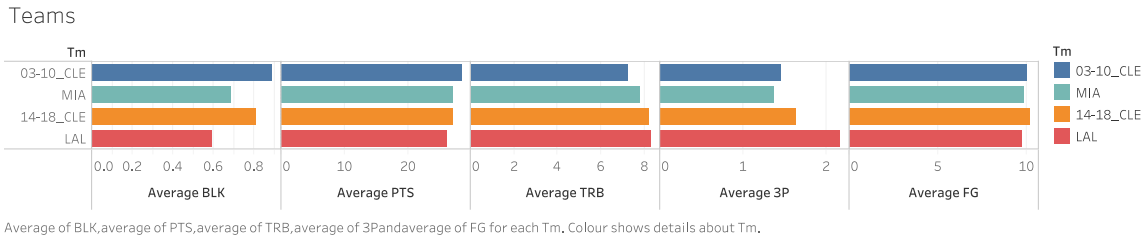
## 4. Data Exploration

The tools I used in this section are Tableau and R.
First of all, I would like to have a overall view of LeBron James yearly stats trend. As shown in Picture 4.1, taking Year as columns and average PTS, BLK, TRB, 3P and FG as rows [appx. D]. And the reason I chose to use average stats is because the tournament status varies from year to year. Representing the average number can eliminate the problem of the game situation as much as possible.



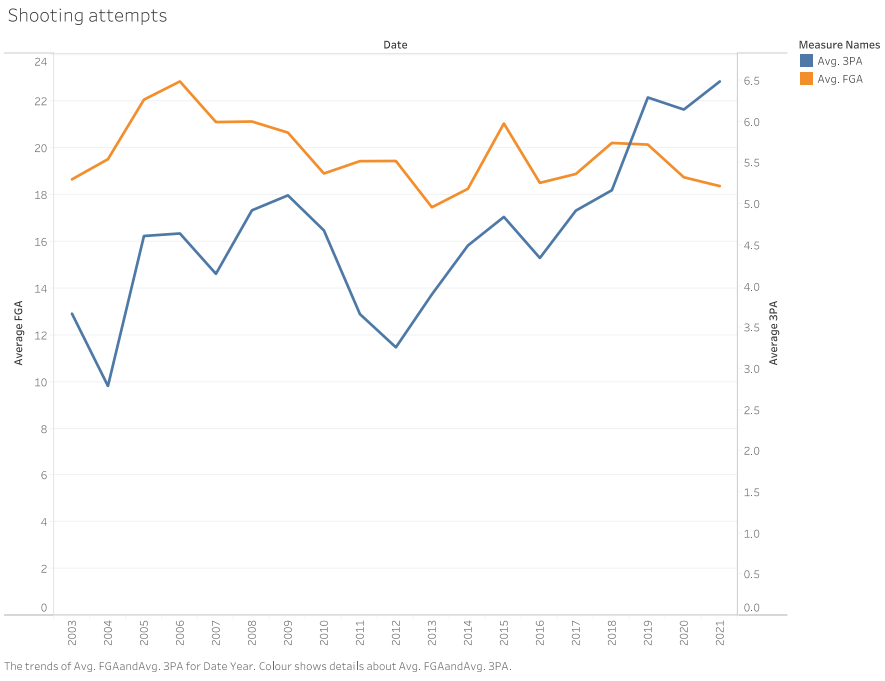
The trends of average of PTS,average of BLK,average of TRB,average of 3Pandaverage of FG for Date Year.
Picture 4.1

Next, I turned years to teams and reverse the graph by changing x to y, y to x. As Picture 4.2, the graph clearly displays how Lebron James's stats changed in different teams.



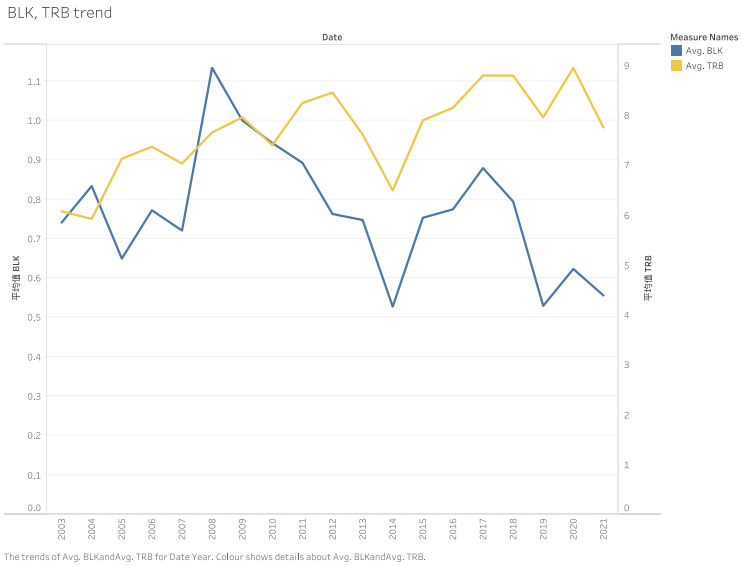

Picture 4.2

Then, offensive method is also one of the references that could understand aging. Putting 3PA, FGA [appx. D] together presented by line chart is the way to view the trend of offensive method change.



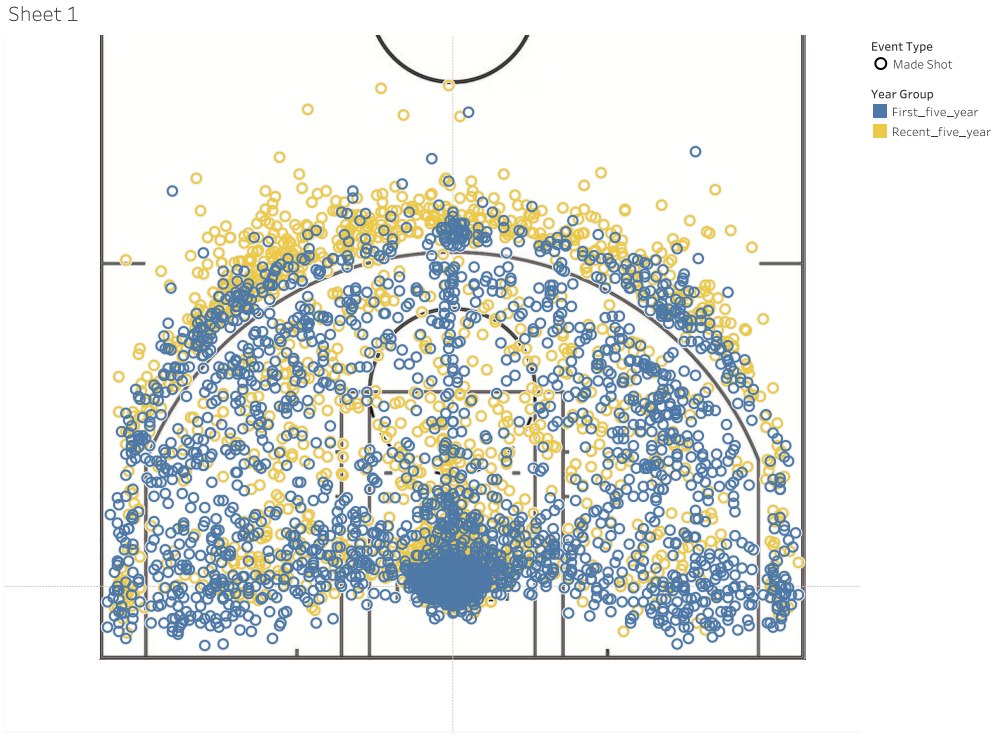

Picture 4.3

As for the trend of BLK and TRB, join these two rows together by clicking "Dual Axis", the trends are shown as below.




Picture 4.4

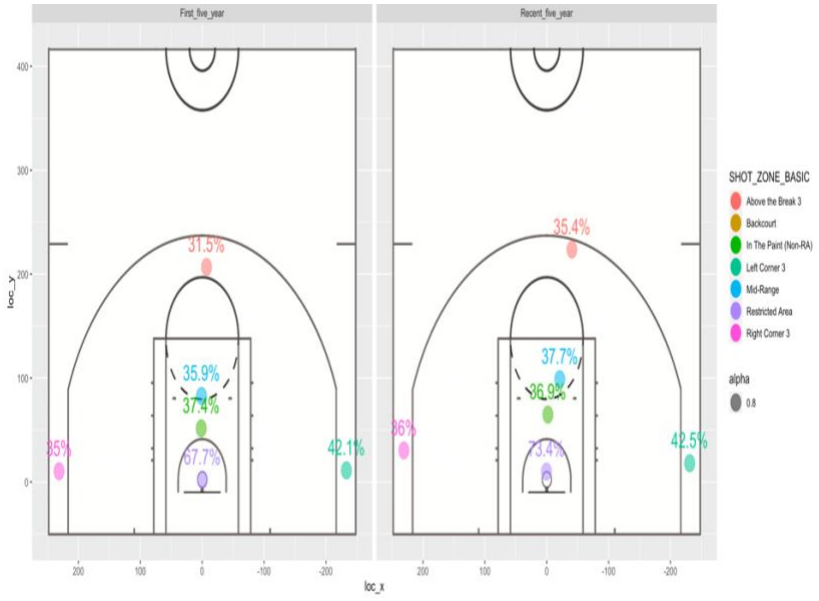
Then the graph [2] [3][4] shown below is the data of shooting, separating by first five years and recent five year in two colors which can tell that how the shooting area change.By using tableau, set location x and y into columns and rows, and import background image [2] and set up the width and length which fit in location range, then take year group data into mark. Here, the data only displayed that LeBron James made the shot.



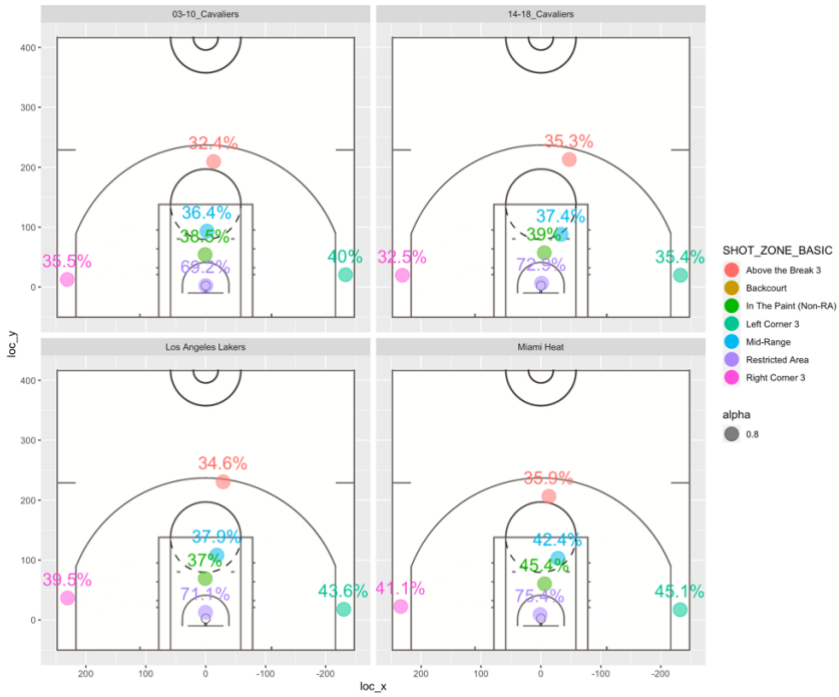

Picture 4.5

Finally, I used R for statistics to calculate the accuracy of shooting in different zone, by summing “made shot” divided the sum of attempt shootings. Different zones are marked with different colors, and also labelled the accuracies of each zone. As for the location of each tag, using mean of each zone’s x and y to decide the tag position.

Picture 4.6 tells the stats in first five year and recent five year. In the following graph Picture 4.7 displays the situation in different teams.




Picture 4.6




Picture 4.7

## 5. Conclusion

The yearly trend (Picture 4.1) shows that LeBron James’s career stats performed stably, there is no obvious change or up and down trend. However, if look at 3PA (3

points attempt) and FPA (Field point attempt) trend (Picture 4.3), from the year 2019, at the age of 34, LeBron James had tried more 3 points shooting attempts than Field goal (2 points) attempts.
Then, the trend of BLK (Block) and TRB (Total Rebound) as Picture 4.4, these two lines' trends are roughly similar; however, average BLK slumped to a level worse than ever, with better rebound numbers than ever. In contrast, blocking requires more physical fitness, which may mean that his body is indeed not as good as before.

As for shooting stats, in Picture 4.5, LeBron James had attempted shooting in wider area in recent five years. In addition, As shown in Picture 4.6, shooting percentages are better than the first five years of his career in almost every position.
In addition, his time with Miami Heat was the pinnacle of his career, with excellent performance in all numbers.

To sum up, even though the stats did not differ clearly, aging may still affect LeBron James, since the stats with more 3 points attempts, less blocks and more peripheral shooting which are the playing style that can decrease body collisions.

## 6. Reflection

There are many different ways to display visualization of sport data, audience of these sport games usually care about offensive performance data, such as field goal percentage, scoring position, scoring hot spots, etc; as a result, the performance with the background court map can better let everyone understand the overall performance.In this project I have learned how to display the visualization with background image; however, I think there may be other players that can perform the data better, such as the players who have grown significantly and after data visualization, it will better reflect the players in the early and late period.
In addition, even though LeBron James may dominate the game very often, basketball still a team work. So, the next time I may try to analyze with more factors, such as the stats compared with different main players or how the teammates impact the stats and so on.

## 7. Bibliography

[1] Owen Phillips (2020), *How to download NBA shot data with R,* URL: https://www.owenlhjphillips.com/new-blog/2020/6/11/how-to-download-nba-shot-data-with-r

[2] Robert Carroll (2016), *Tableau XY Data Plot*, URL:https://www.youtube.com/watch?v=duMB6gaKyvQ

[3] Ed Maia (2015), *How to create NBA shot charts in R*, URL: https://thedatagame.com.au/2015/10/07/visualising-nba-shot-charts-in-tableau/

[4] Daniel Teo (2020), *NBA Shot Charts Part 2: Building the viz in Tableau,* URL: https://datavizardry.com/2020/02/03/nba-shot-charts-part-2/

# Appendix

A. LeBron James's stats in regular session and playoff games from 2003 to 2021
Whole links:
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2004
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2005
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2006
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2007
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2008
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2009
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2010
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2011
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2012
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2013
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2014
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2015
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2016
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2017
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2018
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2019
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2020
https://www.basketball-reference.com/players/j/jamesle01/gamelog/2021

B. LeBron James's shooting distance in regular session and playoff games from 2003 to 2021
Whole links:
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2003-04&PlayerID=2544&ContextMeasure=FGA&Season=2003-04&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2004-05&PlayerID=2544&ContextMeasure=FGA&Season=2004-05&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2005-06&PlayerID=2544&ContextMeasure=FGA&Season=2005-06&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2006-07&PlayerID=2544&ContextMeasure=FGA&Season=2006-07&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2007-08&PlayerID=2544&ContextMeasure=FGA&Season=2007-08&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2008-09&PlayerID=2544&ContextMeasure=FGA&Season=2008-09&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2009-10&PlayerID=2544&ContextMeasure=FGA&Season=2009-10&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2010-11&PlayerID=2544&ContextMeasure=FGA&Season=2010-11&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2011-12&PlayerID=2544&ContextMeasure=FGA&Season=2011-12&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2012-13&PlayerID=2544&ContextMeasure=FGA&Season=2012-13&section=player&sct=plot

https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2013-14&PlayerID=2544&ContextMeasure=FGA&Season=2013-14&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2014-15&PlayerID=2544&ContextMeasure=FGA&Season=2014-15&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2015-16&PlayerID=2544&ContextMeasure=FGA&Season=2015-16&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2016-17&PlayerID=2544&ContextMeasure=FGA&Season=2016-17&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2017-18&PlayerID=2544&ContextMeasure=FGA&Season=2017-18&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2019-20&PlayerID=2544&ContextMeasure=FGA&Season=2019-20&section=player&sct=plot
https://www.nba.com/stats/events/?flag=3&CFID=33&CFPARAMS=2020-21&PlayerID=2544&ContextMeasure=FGA&Season=2020-21&section=player&sct=plot

C. Season

NBA regular session start from October and end in April, and playoff session start from April and end in June in every year.
As a result, the career year only counted from October to April or June in the next year.

D. Glossary

**Rk** -- Rank
**G** -- Season Game
**Age** -- Player's age on February 1 of the season
**Tm** -- Team
**Opp** -- Opponent
**GS** -- Games Started
**MP** -- Minutes Played
**FG** -- Field Goals
**FGA** -- Field Goal Attempts
**FG%** -- Field Goal Percentage
**3P** -- 3-Point Field Goals
**3PA** -- 3-Point Field Goal Attempts
**3P%** -- 3-Point Field Goal Percentage
**FT** -- Free Throws
**FTA** -- Free Throw Attempts
**FT%** -- Free Throw Percentage
**ORB** -- Offensive Rebounds
**DRB** -- Defensive Rebounds
**TRB** -- Total Rebounds
**AST** -- Assists
**STL** -- Steals
**BLK** -- Blocks
**TOV** -- Turnovers
**PF** -- Personal Fouls
**PTS** -- Points
**GmSc** -- Game Score
**+/-** -- Plus/Minus