Analysis of baseball data for performance measure and prediction

By Anirudh K M and Naren Suri

Abstract

The main aim of the project would be to analyze the baseball dataset to understand the team performance and predict their performance in the near future.

The spirit of game is increasing by manifolds these days. The competition between teams is intense and selecting the best team from the squad is like a cherry picking, and this involves a strategy with analytics and statistics to back the decision.

The aim of this project is to analyze the baseball data and understand the performance of the each individual over various performance metrics like, batting, pitching and fielding performances. From these metrics the contribution of each player to the team can be evaluated. Also, using these metrics the future performance of the player can be predicted.

What is Sabermetrics?

Sabermetrics is defined to be the empirical analysis done in the baseball which involves the statistics to measure various parameters in the game. The term sabermetrics is derived from the acronym **SABR,** which means Society for American Baseball Research. This term was coined by Bill James who is considered as one of the big faces in the sabermetrics domain.

Why Sabermetrics?

There are numerous reasons for teams to adopt the techniques in sabermetrics which as follows

- The main goal of sabermetrics is to help the team win more matches.
- The player's performance can also be measured, such as what type of pitches a particular pitcher is expected to deliver. The zone where a batsman is expected to hit the pitch.
- Analyses such as how a batman plays a particular pitcher.
- The fitness of the players can be accessed by these techniques.
- Apart from the players, the performance of the managers can be analyzed with these techniques too.

Data for Sabermetrics

In order to performance these kind of analyses we need data, the data that are used are as follows:

- **Lahman's Data**: This data gives us the details of the each player based on batting, pitching, fielding. Also we have data for the post season, teams, manager's. Each year it is being updated and maintained well by journalist Sean Lahman.

- **Retrosheet data**: This data gives us event by event action that took place in all the matches, it records the batters, pitchers, players in all fielding position, players in other bases and many more in depth details.
- **PITCHfx data**: This data also provides us the pitch by pitch data, where we will access to speed of pitch, other parameters such as velocity, acceleration, spin on the pitch and more.

Software used:

The main software that is used in our analyses is R. The main reasons to use R is as follows:

- One of the best tools for statistical analysis.
- Visualization of data can be made with greater flexibility.
- It is platform independent and it is an open source software.
- It is much easier to use compared to other programming languages.

We used plotly and ggplot for visualization of the data, plotly is proved to be one of the best tools for data visualization. We were able to generate 3D plots and 2D interactive plots with plotly, that is the reason why we opted for that.

How performance is being measured?

The performance of the players can be analyzed through various metrics. The players are mainly classified as

- Batter
- Pitcher
- Fielder

So we have different metrics to measure the performance of player based on these.

Batting metrics

Here are few of the important batting metrics

- **1B** – Single: These are hits which the batsman reaches the first base without any fielding error.
- **2B** – Double: These are hits which the batsman reaches the second base without any fielding error.
- **3B** – Triple: These are hits which the batsman reaches the third base without any fielding error.
- **At bats**-AB: This is the number of plate appearance which does not include base on balls, being hit by pitch, sacrifices and obstruction.
- **Hits**-H: This is the number of times the batter reached any base while batting without any error committed by the defense.
- **Batting Average**-BA: This is calculated as hits divided by at-bats.

- **Base on balls**-BB: A batsman is awarded a walk when the pitcher pitches the pitch outside the strike zone for four times, due to this the batter is awarded first base.
- **Home Run**-HR: These are hits where the batter touches all the four bases successfully without any fielding error.
- **Strike out**-SO: A batter is given a strike out when he swings and misses the pitch third time when the pitch is delivered in the strike zone.
- **On Base Percentage**-OBP: It is defined as the measure of how many times the batter reaches any base without any fielding error, obstruction and interference.
- **Slugging Average**-SLG: It is a statistic used to measure the power of a batter.
- **Sacrifice Fly**-SF: This happen when the batter hits the pitch in air which leads to lose his wicket, but helps in the base runners to advance.
- **Sacrifice Hit**-SH: This is similar to SF, but here the pitch being hit on the ground.
- **Stolen base**-SB: It is the number of the bases advanced by the runner when the pitch is in the defense team control.
- **Caught Stealing**-CS: This is the number of times the runner is caught while trying to steal a base.

Pitching Metrics

- **Complete Game**-CG: Number of times when only one pitcher was used through out the game.
- **Earned Run**-ER: The number of runs given without any errors.
- **Earned Run Average**-ERA: It is calculated by multiplying ER with 9 and dividing with innings pitched.
- **Fielding Independent Pitching**-FIP: It is measure similar to ERA, expect that the events which are controlled by the pitcher alone is taken into account.
- **Opposition Batting average**-OBA: This statistics shows the ability of pitcher to prevent hits.
- **Shutout**-SHO: The number of games where the pitcher does not concede even a run.
- **Save**-SV: This is the number of times where the pitcher finishes the game without surrendering the lead.
- **Wins**-W: It is number of games where the pitcher was pitching while his team is in the lead and finally wins the game.

Fielding Metrics

- **Assists**-A: It is the number of outs recorded where the fielder has assisted other players especially in the base to get an out.
- **Errors**-E: The number of times where the fielder should have made the save, instead makes an error which results in an advantage for the offensive side.
- **Putouts**-PO: It is number of times where a fielder tags the runner and as a results he is given out.

Analyzing team's performance

The data about the overall team's performance for each year can be found in the file **teams.csv** in the Lahman's data.

Let us take two teams, **Boston Red Sox** and **New York Yankees** and analyze their performance factors for the past hundred years.

The program file **team_analysis.R** computes our desired analysis.

Graphical representation of Boston Red Sox and New York Yankees performance

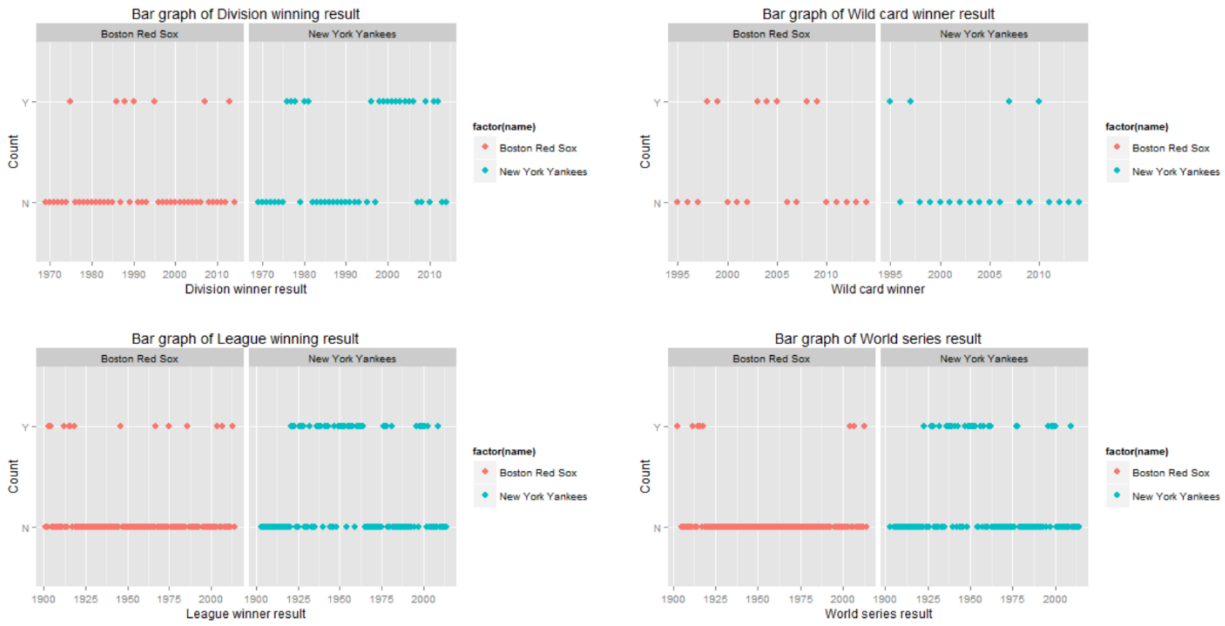


Fig 1. This figure shows the amount of wins BOS and NYA have across in division, wild card, league and world series from early 1900's.

From this figure we can easily see that NYA has won more titles compared to BOS. If we observe closely we could see that especially during the period 1925-1960, NYA has won most of the league and world series title, so we could say that was one of the golden period for the team.

We see that NYA holds the upper hand we see the whole data set, but when we see the results after the year 2000, we don't see much of difference between these two teams in the case of league and world series, but when we see division winner, we see NYA performs better after 2000.

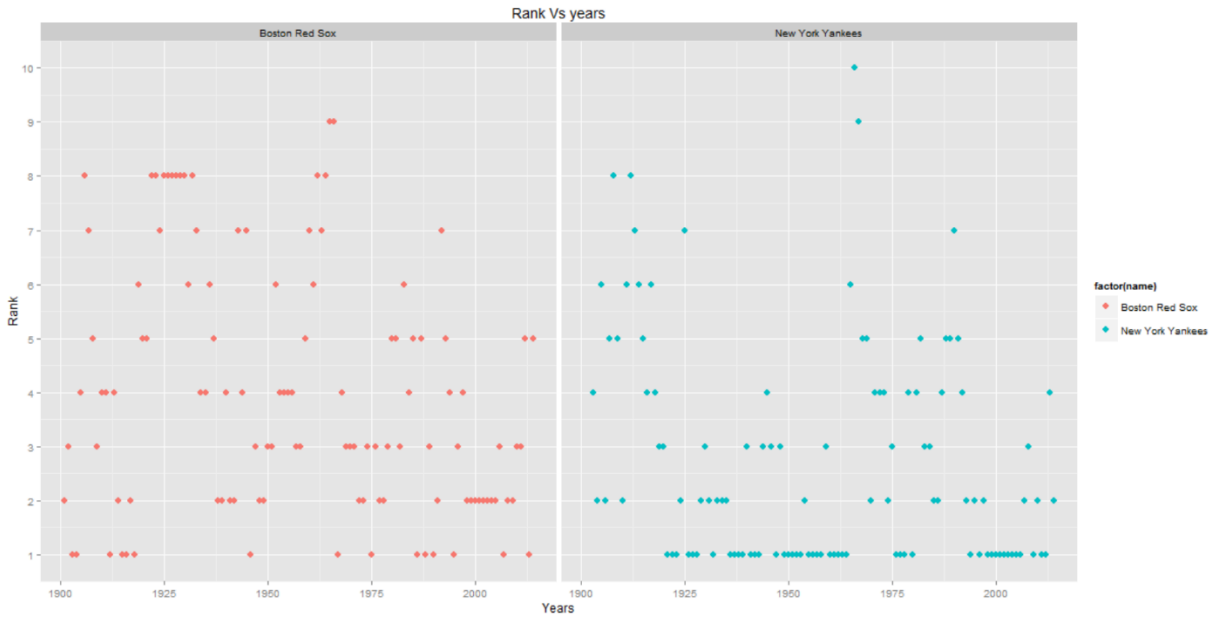


Fig 2. This shows the rank secured by BOS and NYA from the year 1900.

From this figure we can see that most of the times NYA has finished in the top 5 and has secured rank one more times compared to BOS. In this case if we see the performance after 2000, we see that both the teams always end within rank five, this shows that these two teams are top in the current trend.

Fig 3. This graph shows the win-loss ratio of both the teams from the year 1900-2014.

From this figure, in the case of BOS we can see a decline trend in the early part of their career and then we could see a slowly increasing win-loss ratio trend.

In the case of NYA, we could see a constant increase in the win-loss ratio in their early career till 1950. From 1940's till 1975 we could see a decreasing trend for them and finally an increasing trend in the win-loss ratio.

Comparing the two teams, we could easily say that NYA has a better win-loss ratio compared to BOS except in the early part of the career where BOS had better ratio compared to NYA.

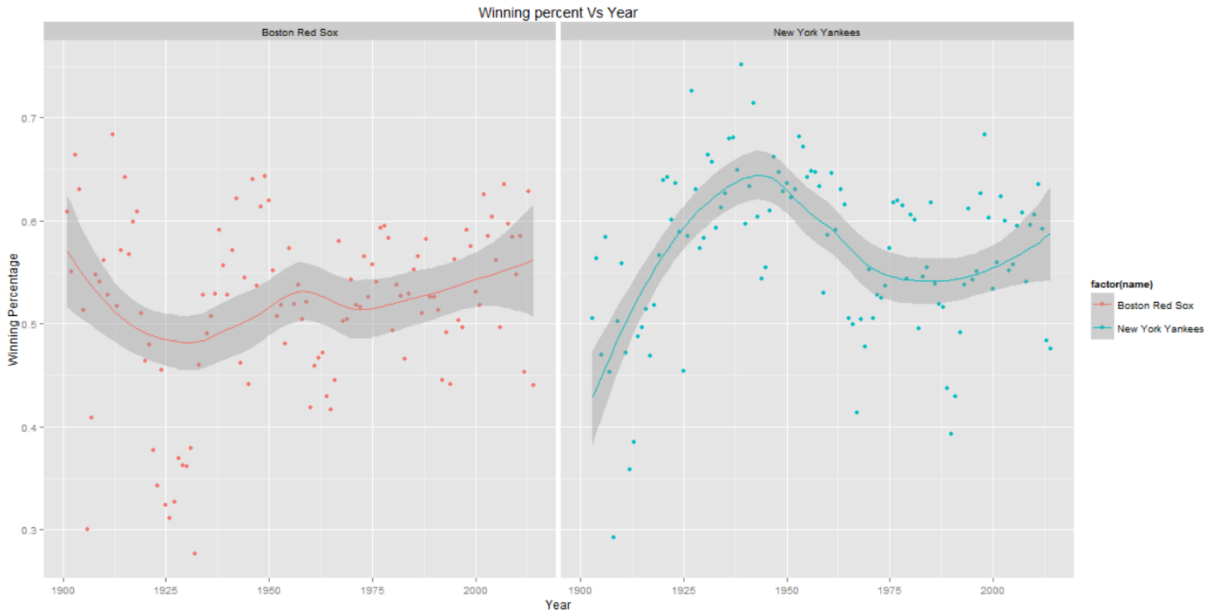


Fig 4. This figure shows the winning percentage value for both the teams from the year 1900.

We can see that the trend in this figure is almost similar to the previous graph (Win-Loss ratio).

The winning percentage in this graph is calculated with the Pythagorean formula rather than the traditional formula.

Fig 5. This graph shows the trend of Hits per atbats and Strikeouts per atbats for both BOS and NYA from the year 1900.

We can observe similar type of trend in both the teams in the figure. With the technology advancement and better batting techniques we could see that the trend in hits per atbats is in the increasing trend from the year 1960.

Also we could see increasing trend in strikeout per atbats as well, this reflects that pitchers have also improved their game and with the help of new pitch types developing, we could see the increase in strike outs as well.

Fig 6. This graph shows the trend in Earned Run Average for both the team from the year 1900.

For the team BOS, we can see that there is an increase in ERA trend through their initial career followed by a dip and a slow increase in trend. We can see similar type of trend for the NYA team also but we with lesser values.

So we would say that NYA has a better ERA compared to BOS, which shows NYA pitchers performance is much better compared to BOS.

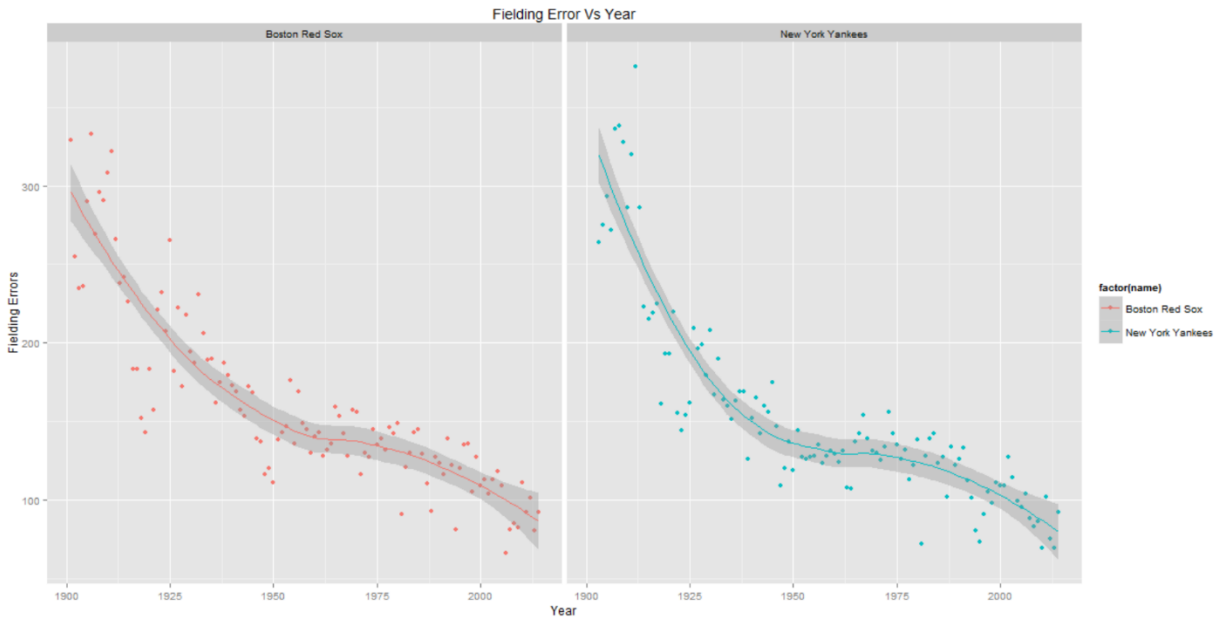


Fig 7. This figure shows the number of fielding error committed by both the teams from the year 1900.

When analyzing the fielding errors we can see that both the teams has improved their fielding skills over the years. So in the case of fielding both the teams proves them they are brilliant.

Conclusions for team performance comparison

So by the analyses of these parameters we can say that NYA has been the better team overall but when we just consider the last five years we can say that both the teams have been going well and surely comes in the best category.

So it will be a tight competition when these two teams face each other.

Batting Performance comparison

The performance of batters can be analyzed with analyzing various batting metrics. For this analysis to be done, we take two players, Albert Pujols and Matt Holliday.

Fig 8. Albert Pujols and Matt Holliday (from left to right)

Albert Pujols: Albert is a right handed batter whose age is 35 years. He currently plays for the Los Angeles Angels.

Matt Holliday: Matt is also a right handed batter whose age is 35 years. He currently plays for the team St. Louis Cardinals.

This analyses is done with both the Lahman's data and the PITCHfx data. The program **batting_lahman.R** uses the lahmans data for analyses and **batting_pitchfx.R** uses the PITCHfx data.

Graphical representation of analysis of PITCHfx data

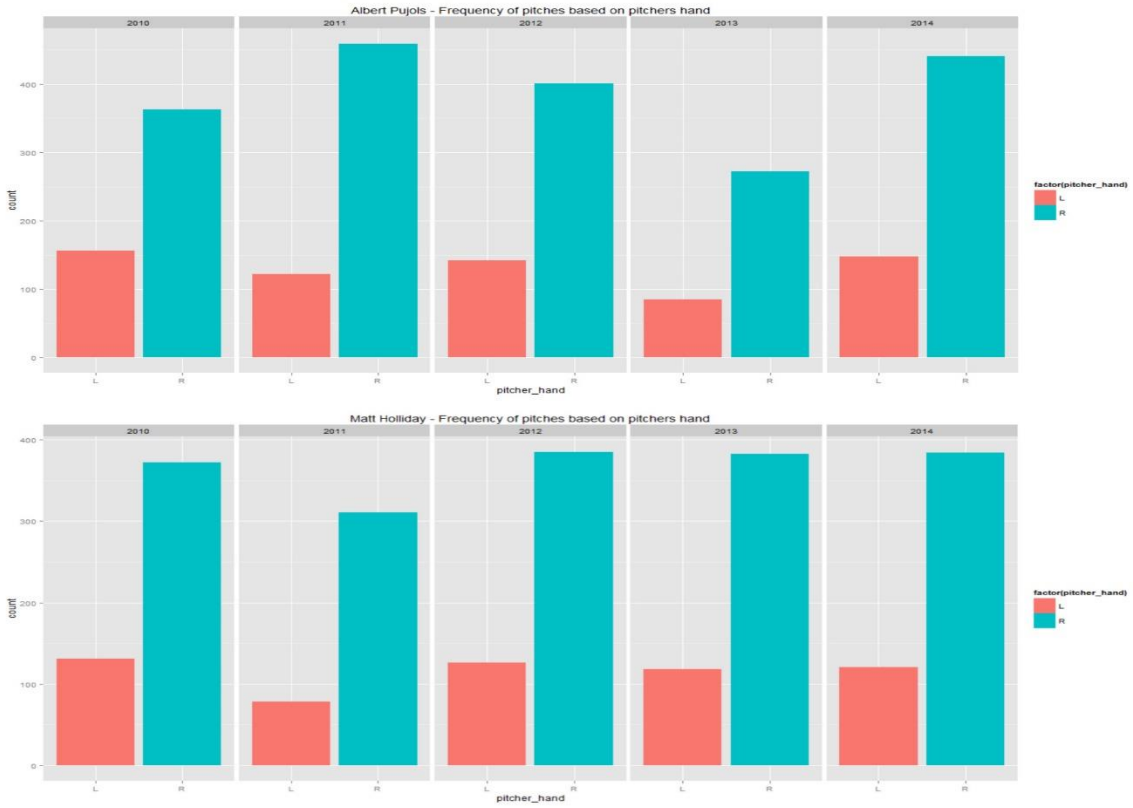


Fig 9. This shows the amount of pitches faced by Pujols and Holliday against right and left handed pitchers from 2010-2014

From the graph we can see that both the batsman has faced more pitches from a right handers compared to left handers.

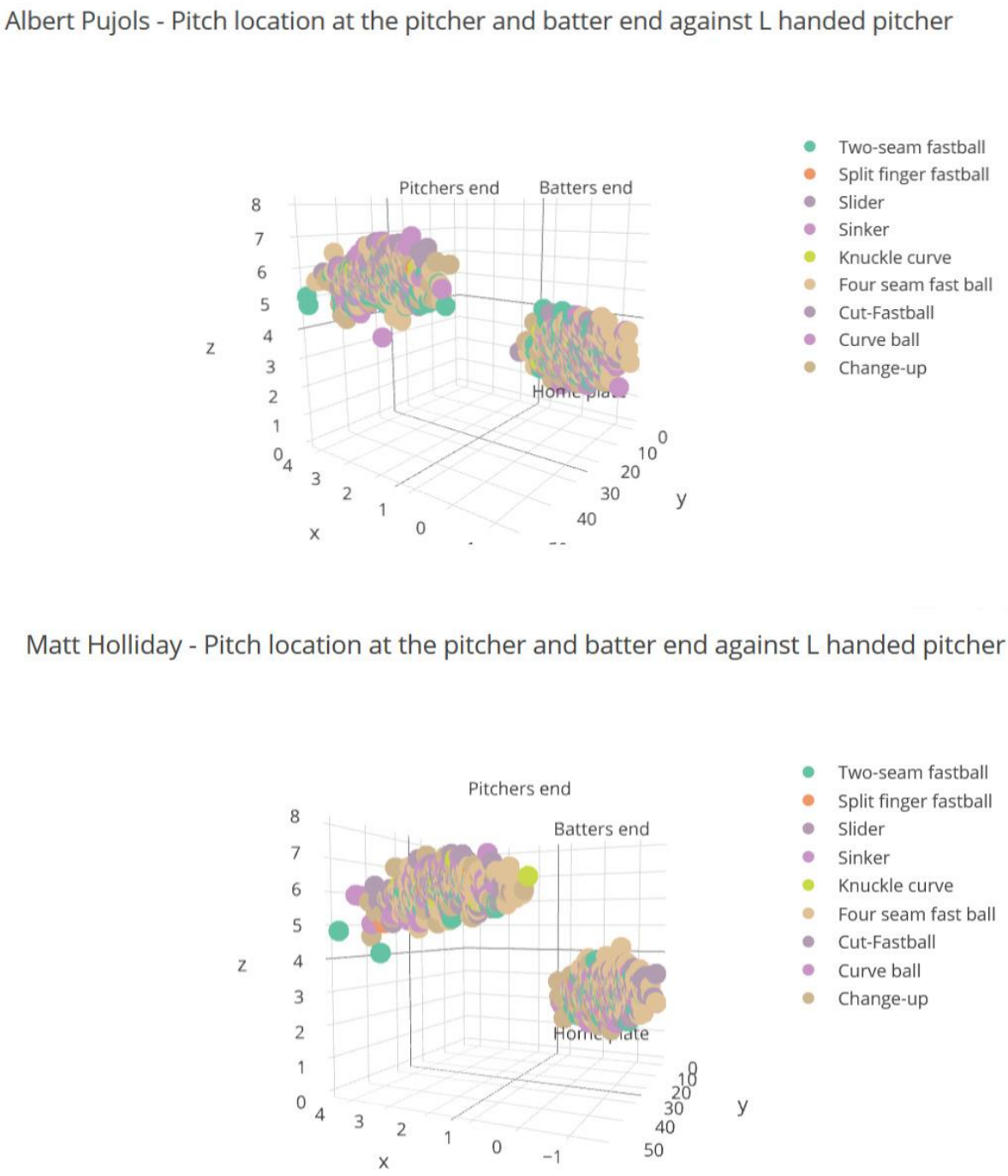


Fig 10. This figure shows the pitch location of pitches at the pitchers and batters end against left handed pitchers for Albert Pujols and Matt Holliday from the year 2010-2014.

To access the 3D features of the image, click here - https://plot.ly/187/~anirudhkm/ - Albert Pujols

To access the 3D features of the image, click here - https://plot.ly/191/~anirudhkm/ - Matt Holliday

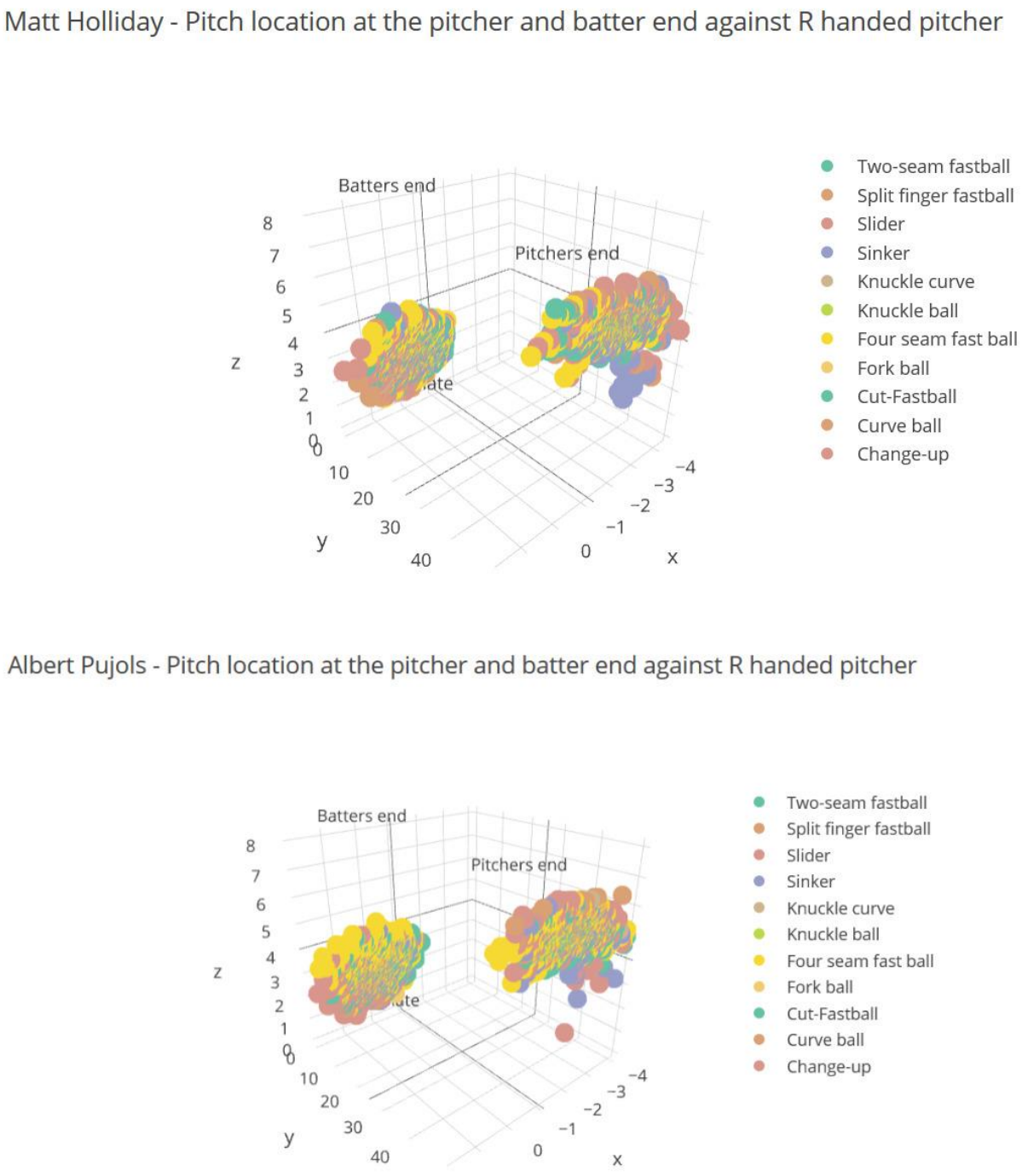


Fig 11. This figure shows the pitch location of pitches at the pitchers and batters end against right handed pitchers for Albert Pujols and Matt Holliday from the year 2010-2014.

To access the 3D features of the image, click here - https://plot.ly/185/~anirudhkm/ - Albert Pujols

To access the 3D features of the image, click here - https://plot.ly/189/~anirudhkm/ - Matt Holliday

In this 3D graph, the x axis corresponds to the right and left side of the home plate for positive and negative value. The y axis corresponds to the distance from the home plate origin. The z axis corresponds to the height of the pitch from the ground level.

From the above graphs we can see the pitch locations at the batters and pitchers end, we can see that the stock ball for both right and left handed pitchers is four seam fastball and slider.

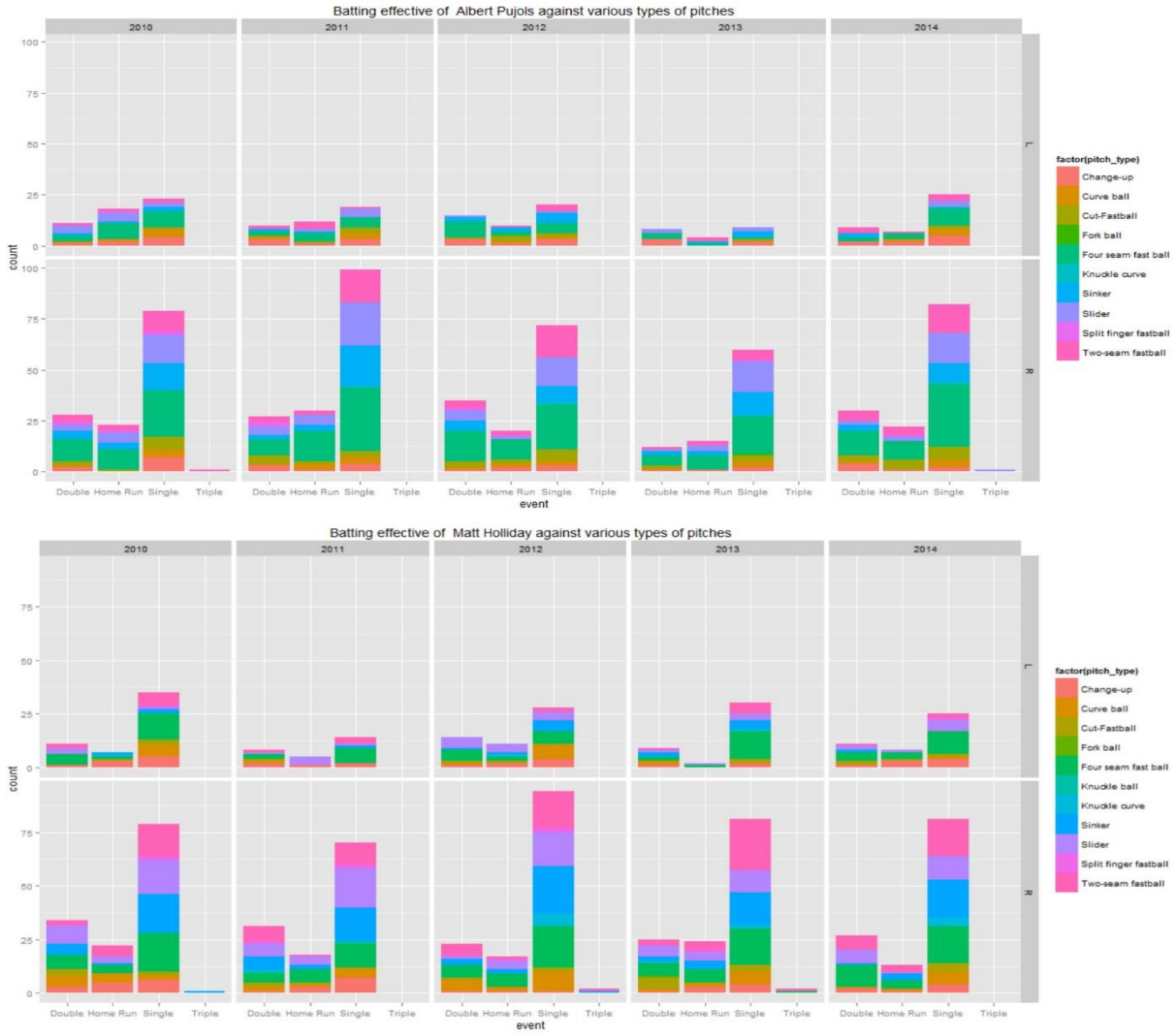


Fig 12. This figure shows the amount of Single, Double, Triple and Home run hit by Albert Pujols and Matt Holliday against various pitch types from the year 2010-2014.

From the above graph we can infer that there is a similar pattern with both these batsman. We see that Albert Pujols has struggled against left hand pitchers compared to Matt Holliday.

Fig 13. This graph shows the pitch location of pitches at the home plate for the events Single, Double, Triple and Home run from the year 2010-2014.

In this graph, positive value for px corresponds to the pitch location to the right side from the middle of the home plate and vice versa. The value of pz corresponds to the height of the pitch in feet from the ground level, so a negative value of pz indicates that the pitch has bounced before reaching the home plate.

In this graph also we could see a similar pattern between both the batsman.

So we can say from the above two graphs that singles are scored more followed by doubles, home run and triples. We can also see that triple scored are being very rare for both the batsman.

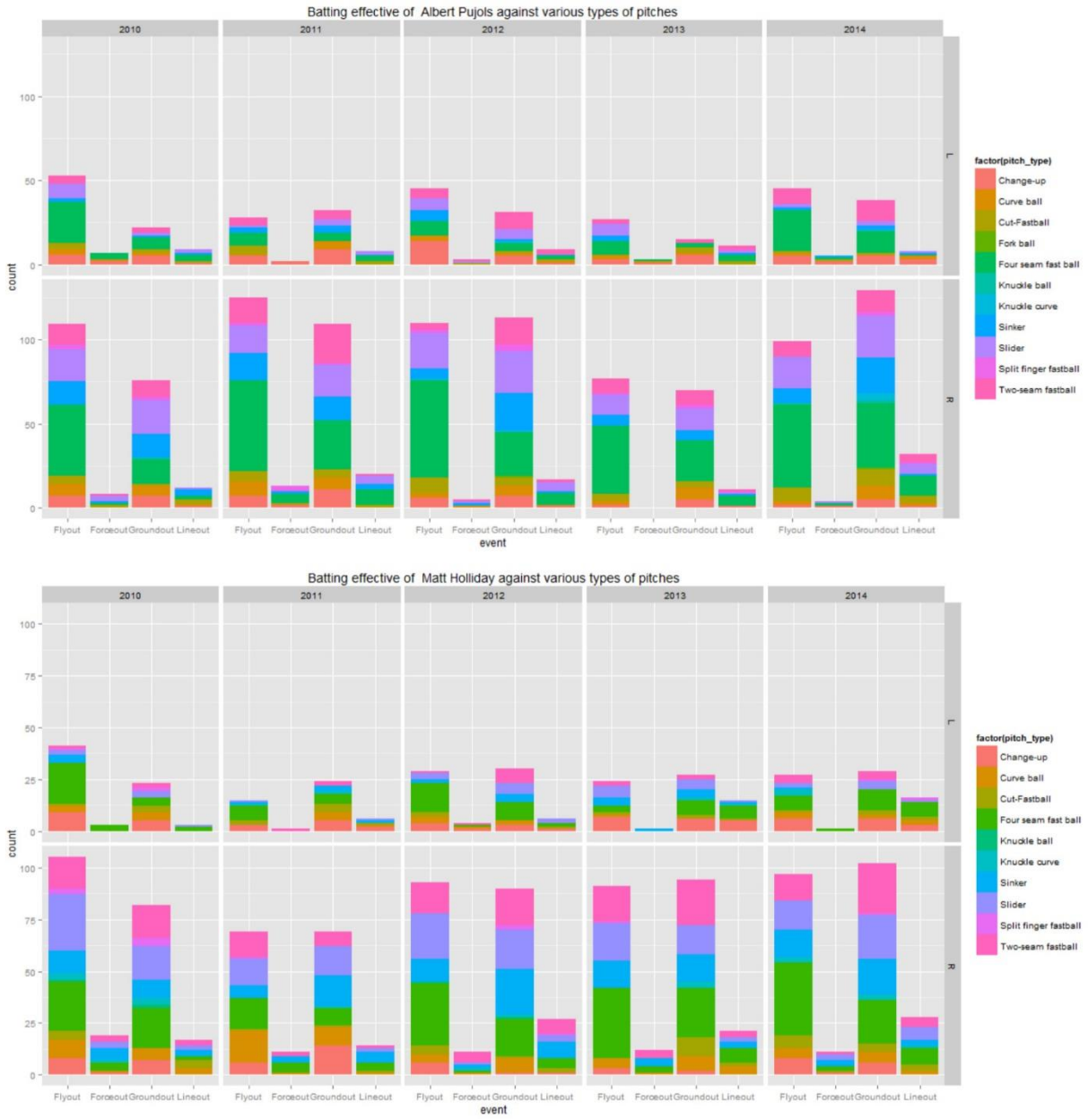


Fig 14. This figure shows the number of flyouts, forceout, groundout and lineout for Albert Pujols and Matt Holliday against various pitch types from the year 2010-2014.

We could see that the both Albert Pujols and Matt Holliday have got out on flyout and groundout majority of the times. When the take into the account of forceout and lineout we see that Matt Holliday has most outs based on that compared to Albert Pujols.

Fig 15. This graph shows the pitch location of pitches at the home plate for the events flyout, forceout, groundout and lineout from the year 2010-2014.

In this graph, positive value for px corresponds to the pitch location to the right side from the middle of the home plate and vice versa. The value of pz corresponds to the height of the pitch in feet from the ground level, so a negative value of pz indicates that the pitch has bounced before reaching the home plate.

In this graph also we could see a similar pattern between both the batsman.

Analysis of batting performance based on Lahman's data

The program batting_lahman.R is used to analyze the batting performance based on Lahman's data.

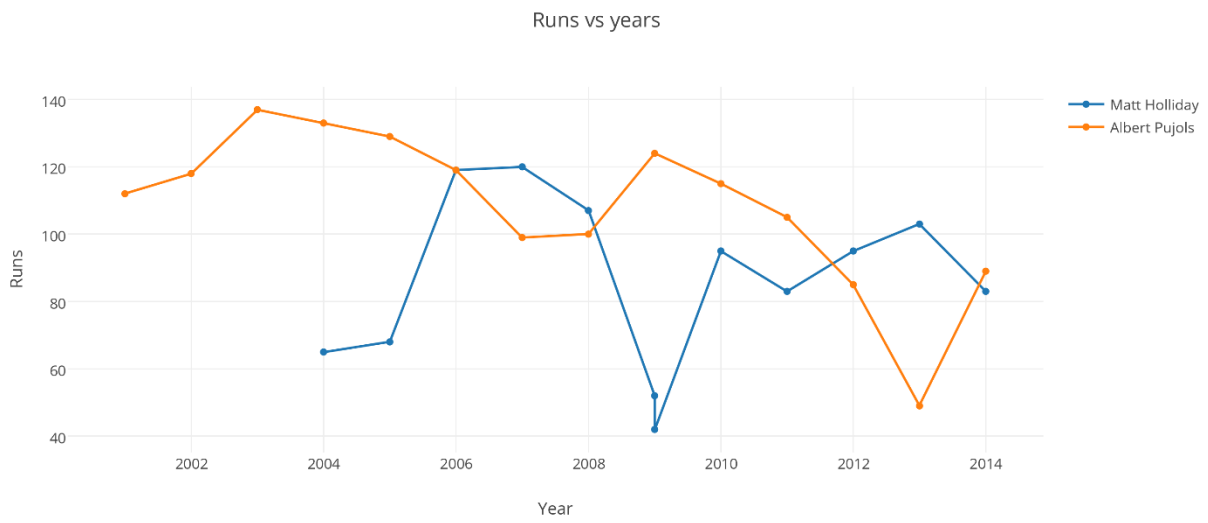


Fig 16. This figure shows the runs scored by Albert Pujols and Matt Holliday through out their career. Complete features of the figure can be accessed - https://plot.ly/100/~anirudhkm/

From this figure we can see that Albert Pujols at his starting year played 161 matches and scored 112 runs. He has been quite consistent throughout all the years. From the year 2009 we could see a dip in his performance and again he has shown sign of improvement, but not at his usual best.

In the case of Matt Holliday, we could see that he had a scratchy start and then played his best during the years 2006-2008. After that we could see a great dip in his run scoring and then he has picked up and maintained a decent run.

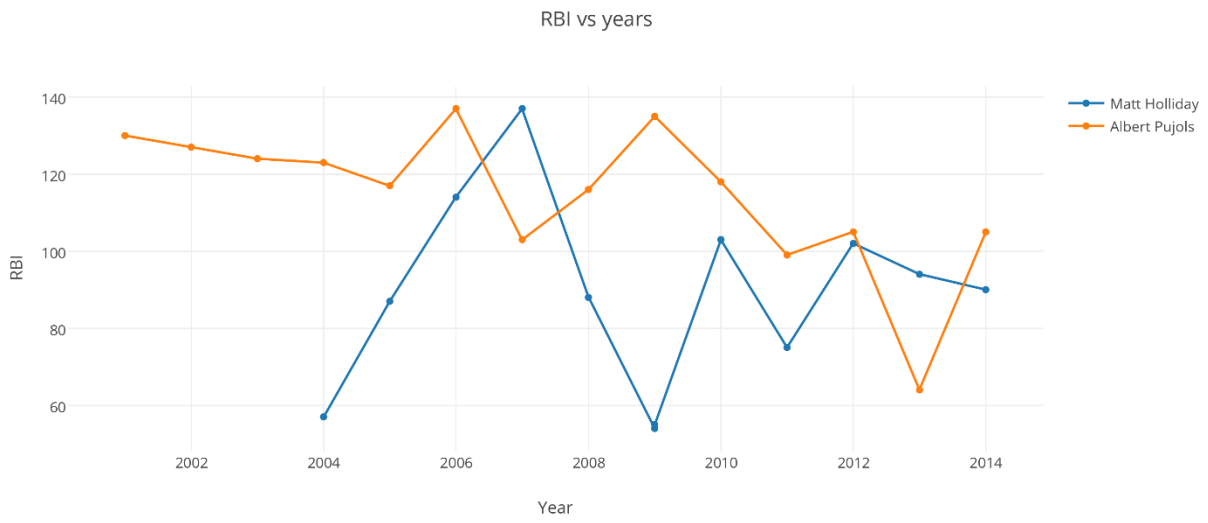


Fig 17. This figure shows the Run Batted In for Albert Pujols and Matt Holliday starting from their career. Complete features of the figure can be accessed - https://plot.ly/102/~anirudhkm/

In the case of RBI, we can see that Albert Pujols has a great RBI throughout his career expect in the years 2007 and 2013.

In the case of Matt Holliday we can see that his RBI count is very inconsistent throughout his career. We can see he peaks very high during the year 2007 only.

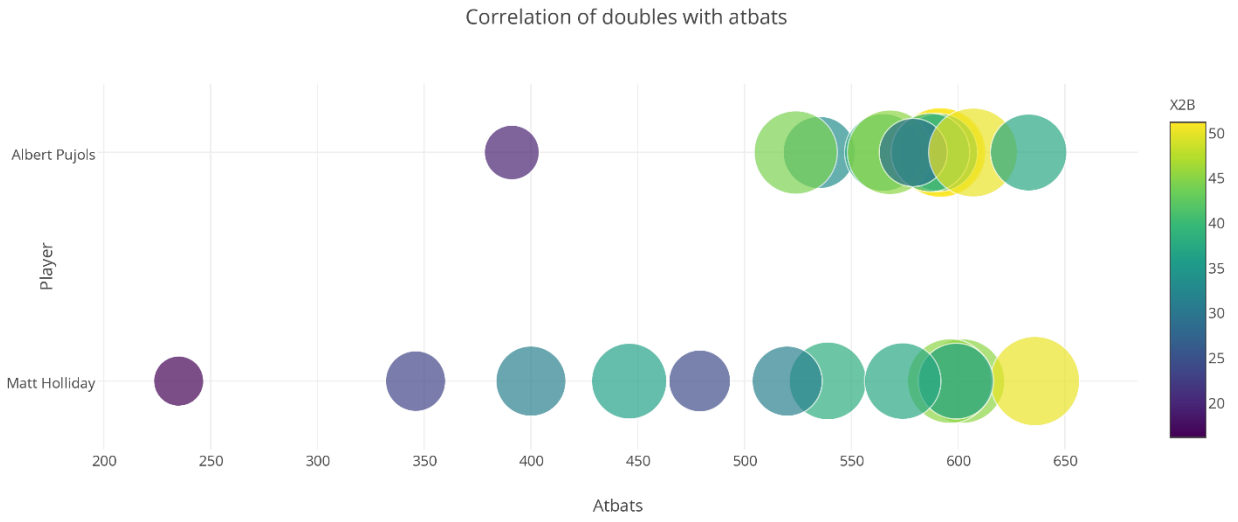


Fig 18. This figure shows the correlation between the amount of atbats and the double for Albert Pujols and Matt Holliday. Complete features of the figure can be accessed here - https://plot.ly/128/~anirudhkm/

From this graph we can infer that for Albert Pujols, we can see that higher the atbats higher the doubles for him. But on the other hand we could see that Matt Holliday doesn't hit doubles as good as Albert Pujols for higher atbats.

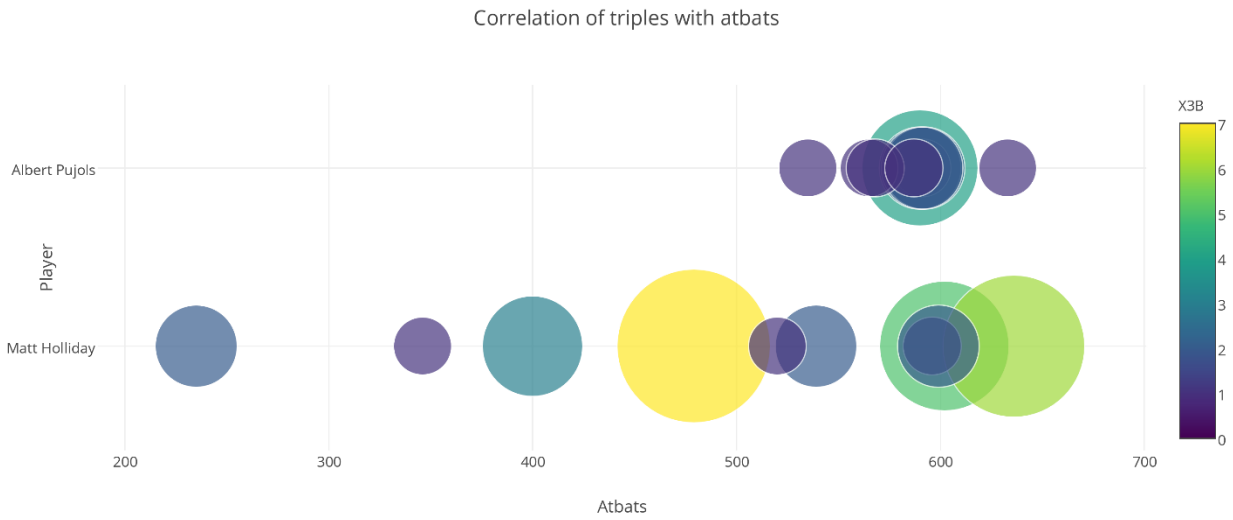


Fig 19. This figure shows the correlation between the amount of atbats and the triples for Albert Pujols and Matt Holliday. Complete features of the figure can be accessed here - https://plot.ly/130/~anirudhkm/

From this figure we could see that Albert Pujols is not good at hitting triples, even for high atbats he his triples scored are very less. But Matt Holliday is better in the case of hitting triples.

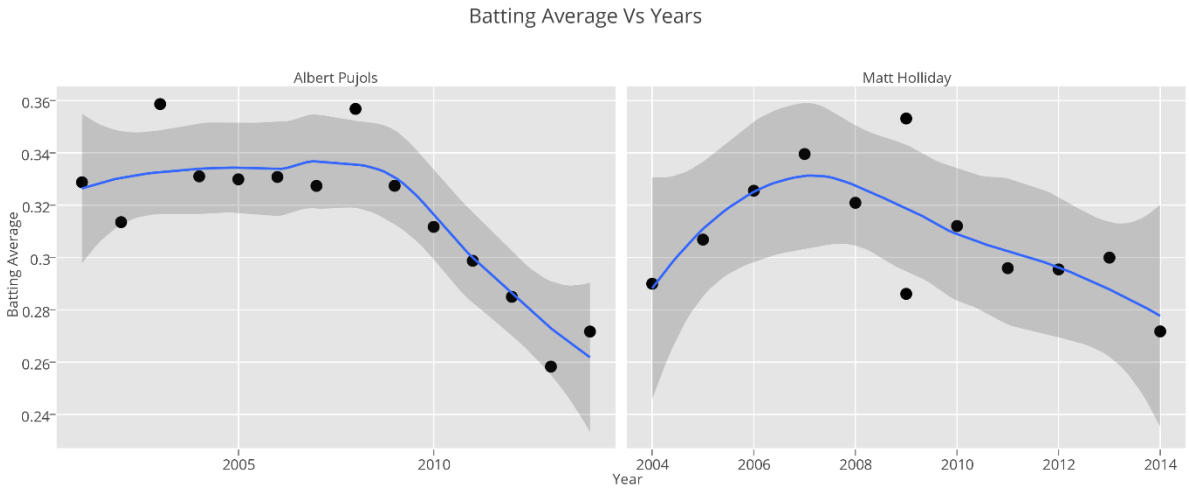


Fig 20. This figure shows the batting average of Albert Pujols and Matt Holliday throughout their career. Complete features of the figure can be accessed here – https://plot.ly/108/~anirudhkm/

From the batting average trend, we could see that Albert Pujols had a constant and higher batting average from his career beginning till the year 2008. After that we could see a decreasing trend in his batting average.

In the case of Matt Holliday, we could see that his batting average has an increase trend and picks at the year 2007 and after that we could see the decreasing trend for Matt too.

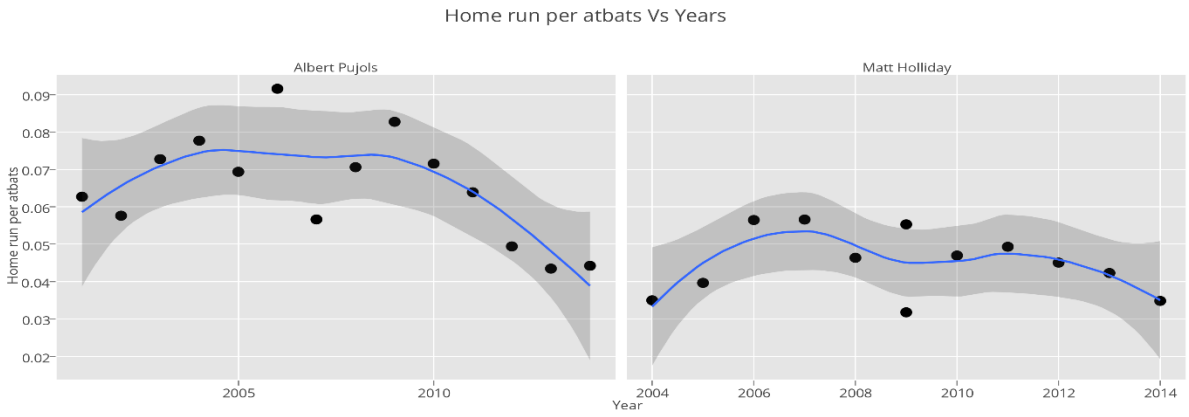


Fig 21. This figure shows the home run per atbats trend for Albert Pujols and Matt Holliday throughout their career. Complete features of the figure can be accessed here - https://plot.ly/110/~anirudhkm/

From the graph we can see that Albert Pujols has shown the ability to hit more HR/AB. But if we observe, we can see that after 2010, there is a decrease in trend. In the case of Matt Holliday, we can see kind steady increase initially followed by a dip and the trend settles down.

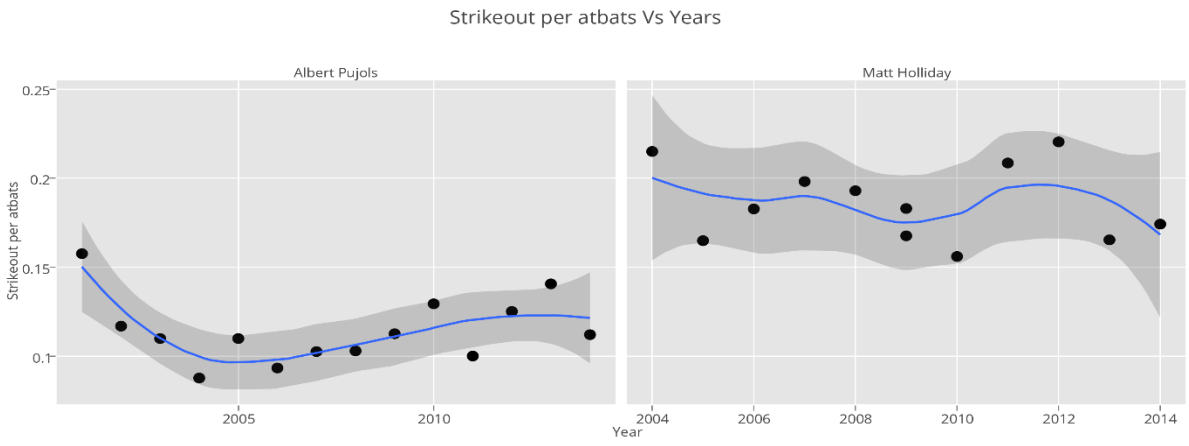


Fig 22. This figure shows the trend of strikeout per atbats for Albert Pujols and Matt Holliday throughout their career. Complete features of the figure can be accessed here - https://plot.ly/112/~anirudhkm/

From the graph we can infer that Albert Pujols has lower SO/AB which shows that he connects the bat with pitch most of the times, on the other hand Matt Holliday has a higher SO/AB which shows that he misses many pitches when compared to Albert Pujols.

So we can say that Albert has better ability to connect the pitch compared to Matt Holliday.

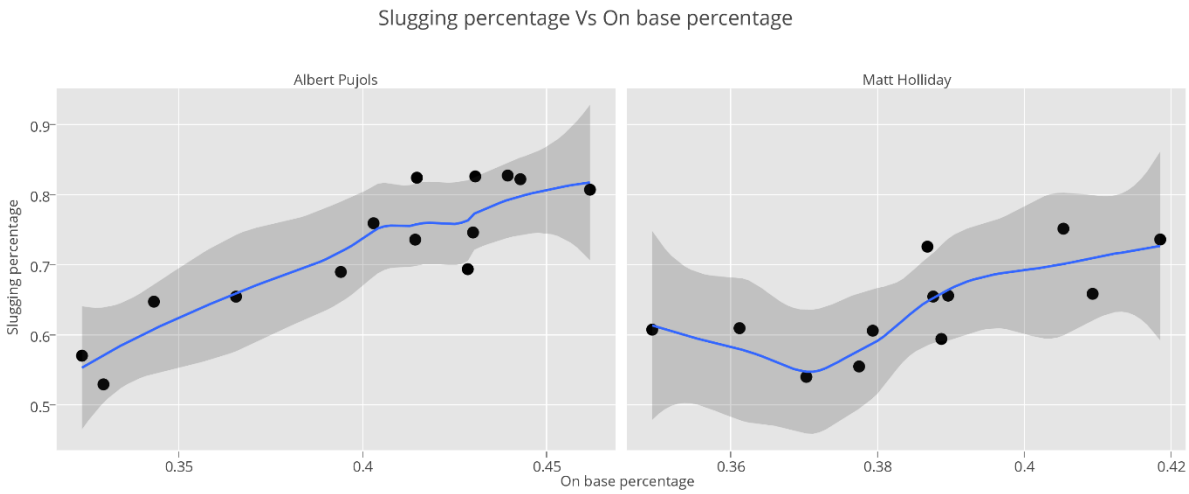


Fig 23. This figure shows the correlation between on base percentage and slugging percentage for Albert Pujols and Matt Holliday. Complete features of the figure can be accessed here - https://plot.ly/116/~anirudhkm/

From this graph we can see that for Albert Pujols, there is a positive correlation between OBP and SLG. But for Matt Holliday we could see a decrease trend followed by an increasing trend with values less than of Albert Pujols.

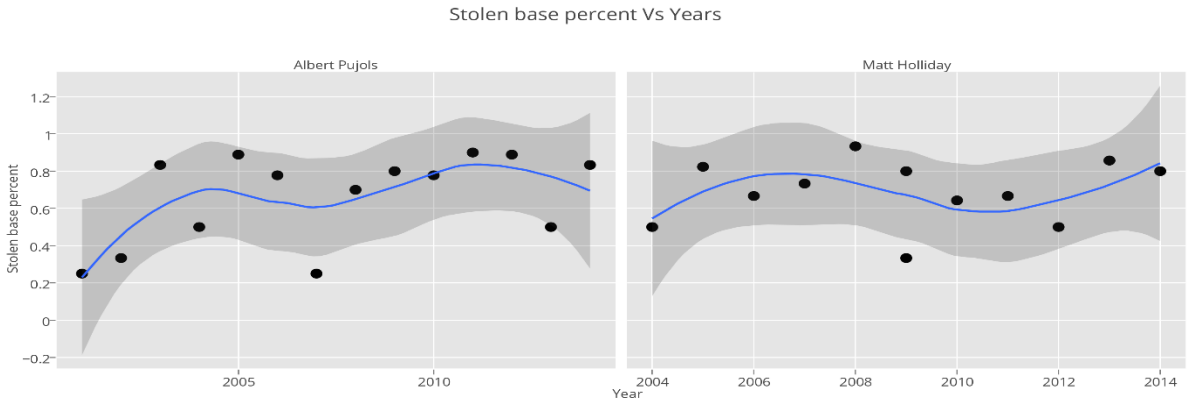


Fig 24. This figure the stolen base percentage for Albert Pujols and Matt Holliday throughout their career. Complete features of the figure can be accessed here - https://plot.ly/118/~anirudhkm/

This figure shows the effective of Albert Pujols and Matt Holliday while they are in other bases. We could see that both of them as a good stolen base percentage. This shows that both of them are equally good when they are in the runner's base.

Fielding analysis

Let us look at how much Albert Pujols and Matt Holliday add value to the team through fielding. The fielding.csv in the Lahman's data set is used to for this analysis. The program which performs these analysis is fielding.R

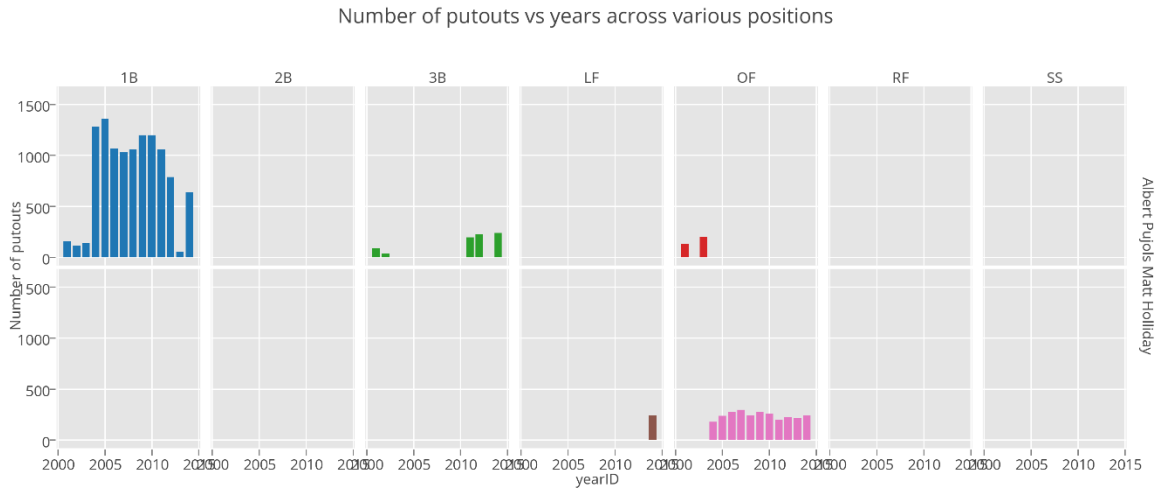


Fig 25. This figure shows the number of putouts while fielding in various positions for Albert Pujols and Matt Holliday throughout their career. Complete features of the figure can be accessed here - https://plot.ly/178/~anirudhkm/

From this graph we can see that Albert has fielding primarily in the 1B position and also in the positions 3B and OF. But he has been achieved more putouts from the 1B position.

On the other hand Matt Holliday has fielding in the OF position most of the times in the LF position too. But he has less number of putouts compared to Albert.

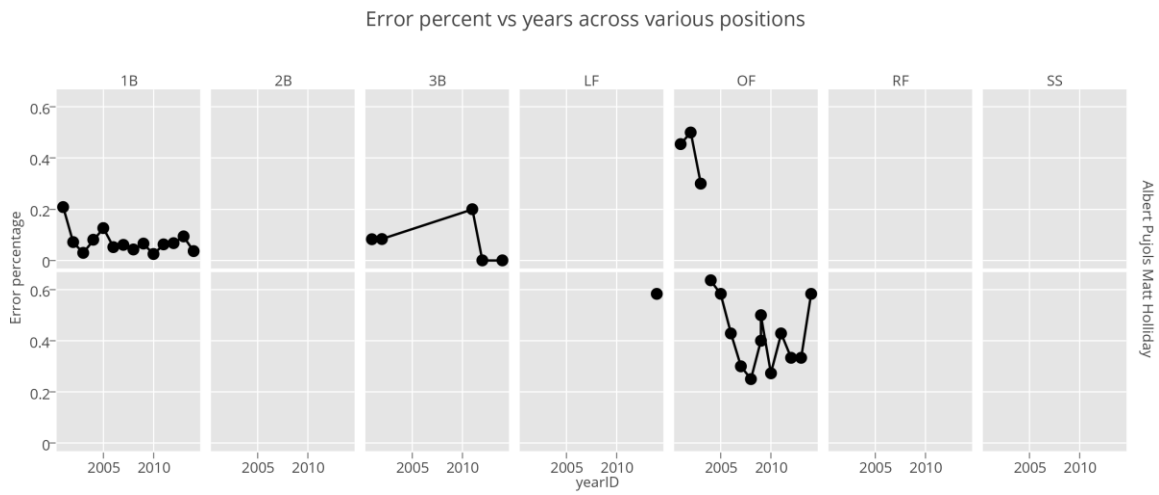


Fig 26. This figure shows the error percentage while fielding in various positions for Albert Pujols and Matt Holliday throughout their career. Complete features of the figure can be accessed here - https://plot.ly/180/~anirudhkm/

From the graph we could clearly observe how good Albert Pujols in the 1B position. When he was fielding the OF and 3B, he seems to commit more errors, which might be reason he wanted to field in the 1B position.

In the Matt Holliday, we can see that he has a higher error percentage compared to Albert Pujols.

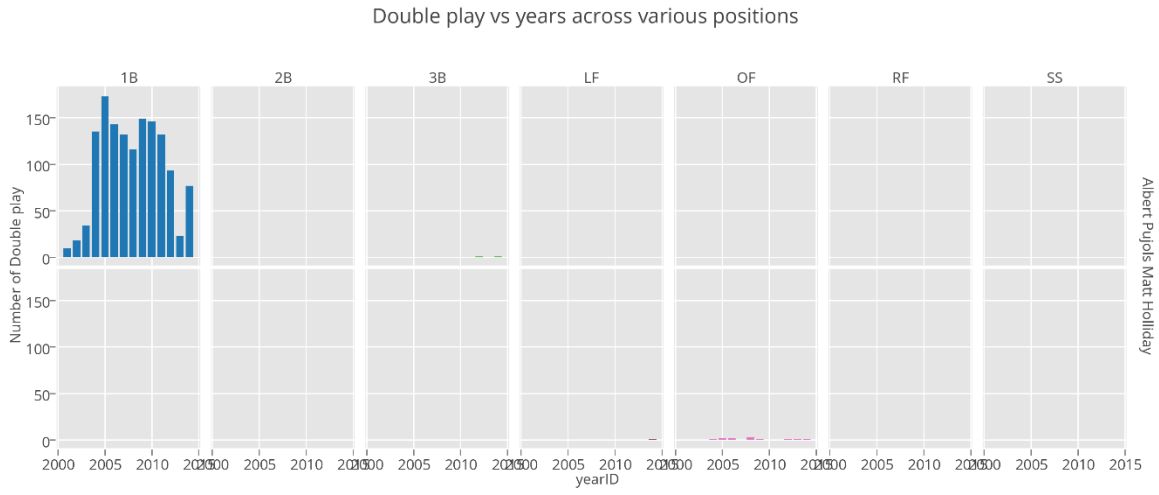


Fig 27. This figure shows number of double plays while fielding in various positions for Albert Pujols and Matt Holliday throughout their career. Complete features of the figure can be accessed here - https://plot.ly/180/~anirudhkm/

Since Albert Pujols fields in the 1B position, he is expected to have more double plays, on the other hand we see little contribution of double play from Matt Holliday too.

Conclusion for batting performance comparisons

So from the various graphs we can say that Albert Pujols was a top notch player till the year 2010. But after 2010 his performance starts to dip, may be because of his age factor. But in the year 2014, he has performed well compared to his performances in the year 2010-2013. So we can have little hope on Albert Pujols to play well in coming seasons. But we could say how good a player he was in his overall career.

In the other hand we could see that Matt Holliday's performance is quite inconsistent, this can be evident from many ups and downs in his performance graph.

When we consider the fielding aspect we can say that Albert seems to be a good fielder in the 1B position, also we could say that Matt Holliday is an ordinary fielder.

So, we can say that overall Albert Pujols is way better batter compared to Matt Holliday. But from the year 2010-2014, we can see that both of them hasn't performed to their best performance that they set.

Pitching analysis

Pitching is one of the most important aspects of a baseball game, if a team has a very good pitcher, then their chances of winning is going to very high.

So we take the pitchers Pat Neshak and Randy Choate for our analysis and compare them.

Fig 28. Pat Neshak and Randy Choate(left to right)

Pat Neshak is a right handed pitcher whose current age is 35 years and he pitches for the team Houston Astros.

Randy Choate is a left handed pitcher whose current age is 40 years and he pitches for the team St. Louis Cardinals.

The data to analyze pitching performances can be taken from Lahman's data and PITCHfx data. The program pitching_lahman.R, pitch_location_3D.R and pitching_pitchfx.R helps to do these analysis.

Pitching analysis with PITCHfx data

In this data, we have the coordinates of the pitch location, for each pitch what type of event that happened along with details such as start and end speed and the pitch type that is being bowled.

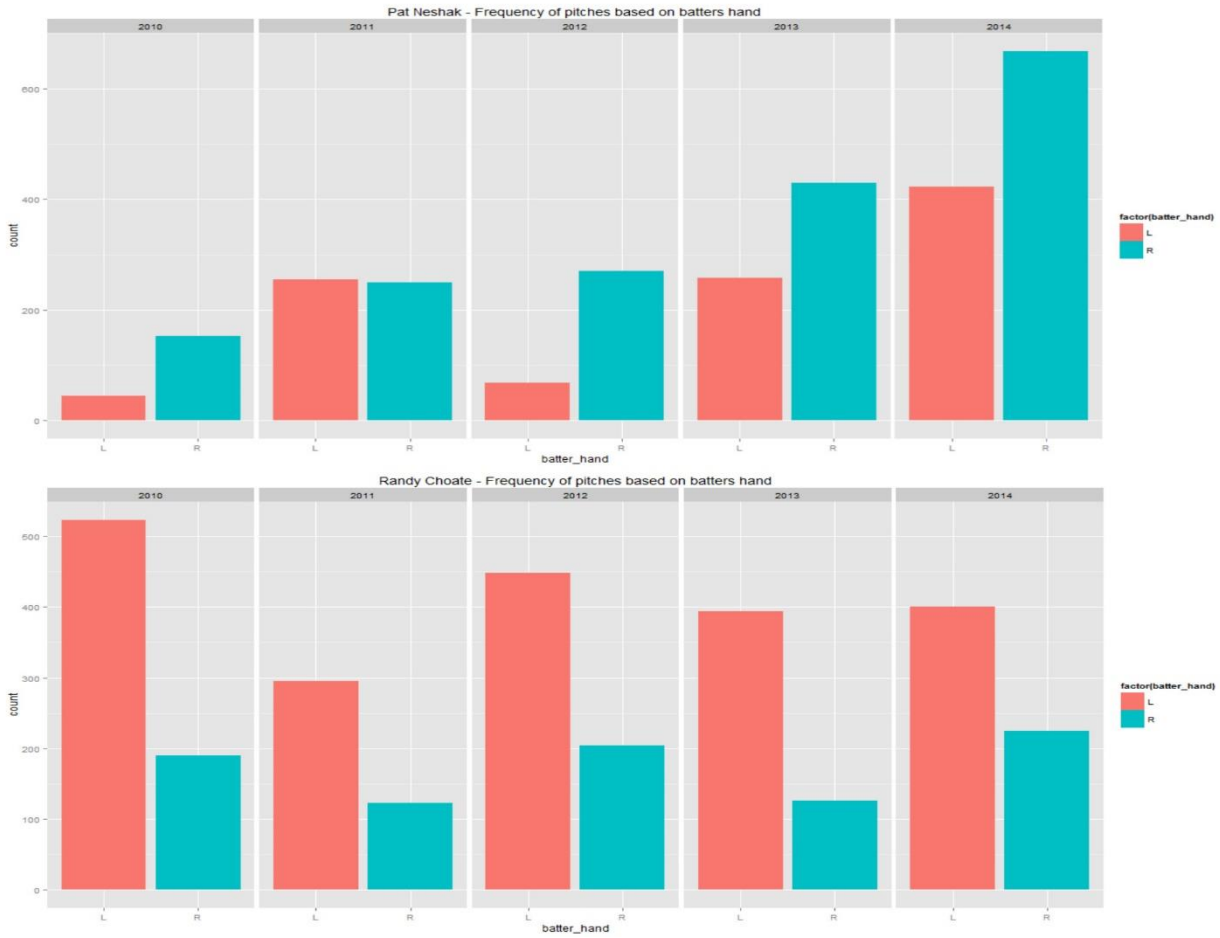


Fig 29. This figure shows the amount of pitches pitched by Pat Neshak and Randy Choate to both left and right handed batter from the year 2010-2014.

From this figure we could see that Pat Neshak has pitched most of the times to right handed batter and whereas Randy Choate has pitched most of his pitches to left handed batters.

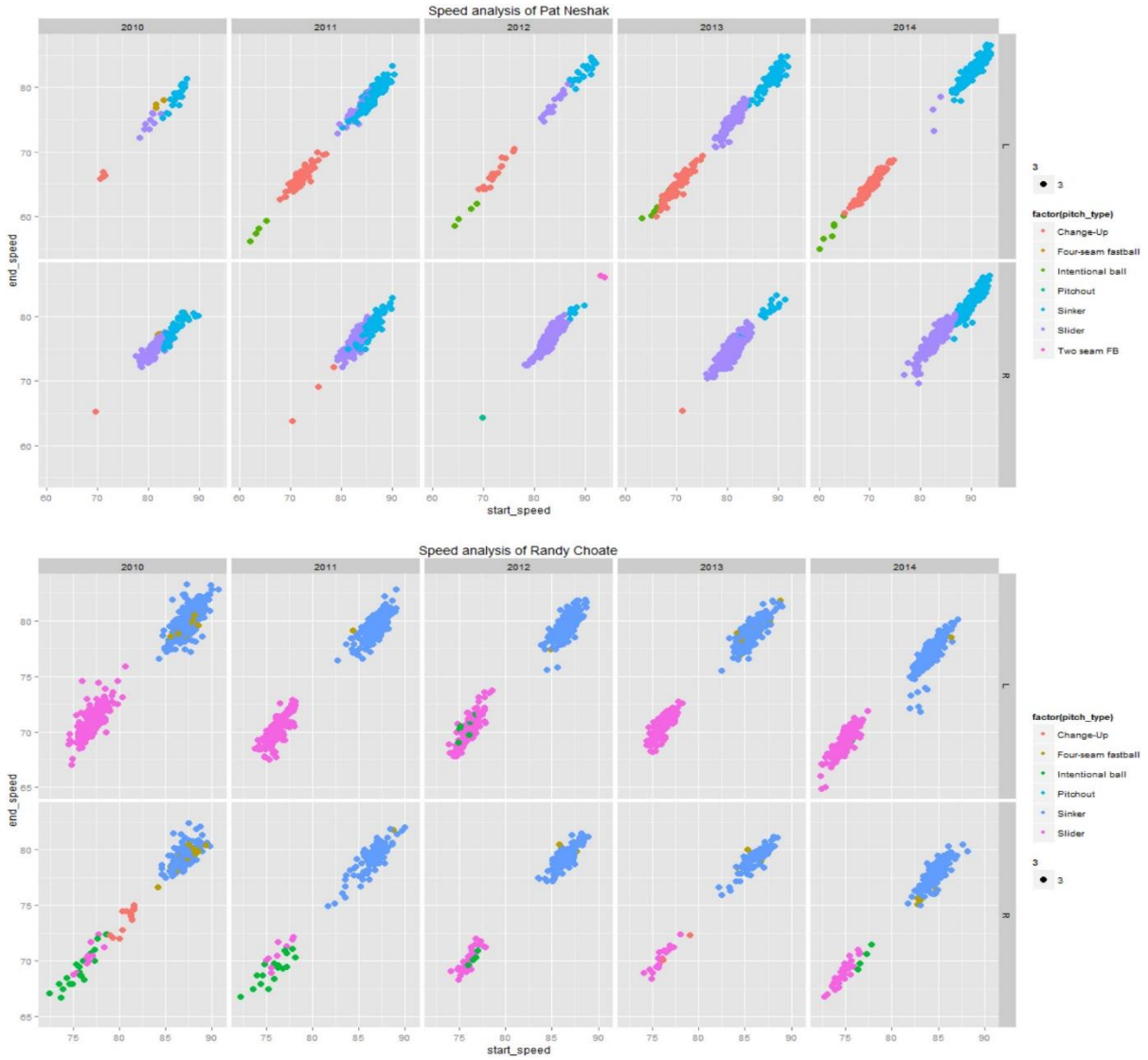


Fig 30. This figure shows the relation between end speed and start speed for Pat Neshak and Randy Choate for various pitch types and batters hand from year 2010-2014.

From this graph we can see that Pat Neshak sinker and change up pitch types to the right handed batter most of the times, but for left handed batters we could see that he sticks with sinkers and sliders majority of the times. So we could say that Pat tries to pitch with more variety to the right handed pitchers.

In case of Randy Choate, we could see that for both right and left handed pitchers, he sticks with the pitch type sinker and slider most of the times.

Another observation that we can make is that we see Pat Neshak being the quicker pitcher when compared both, we could see the difference in speed of slider pitch type to infer this.

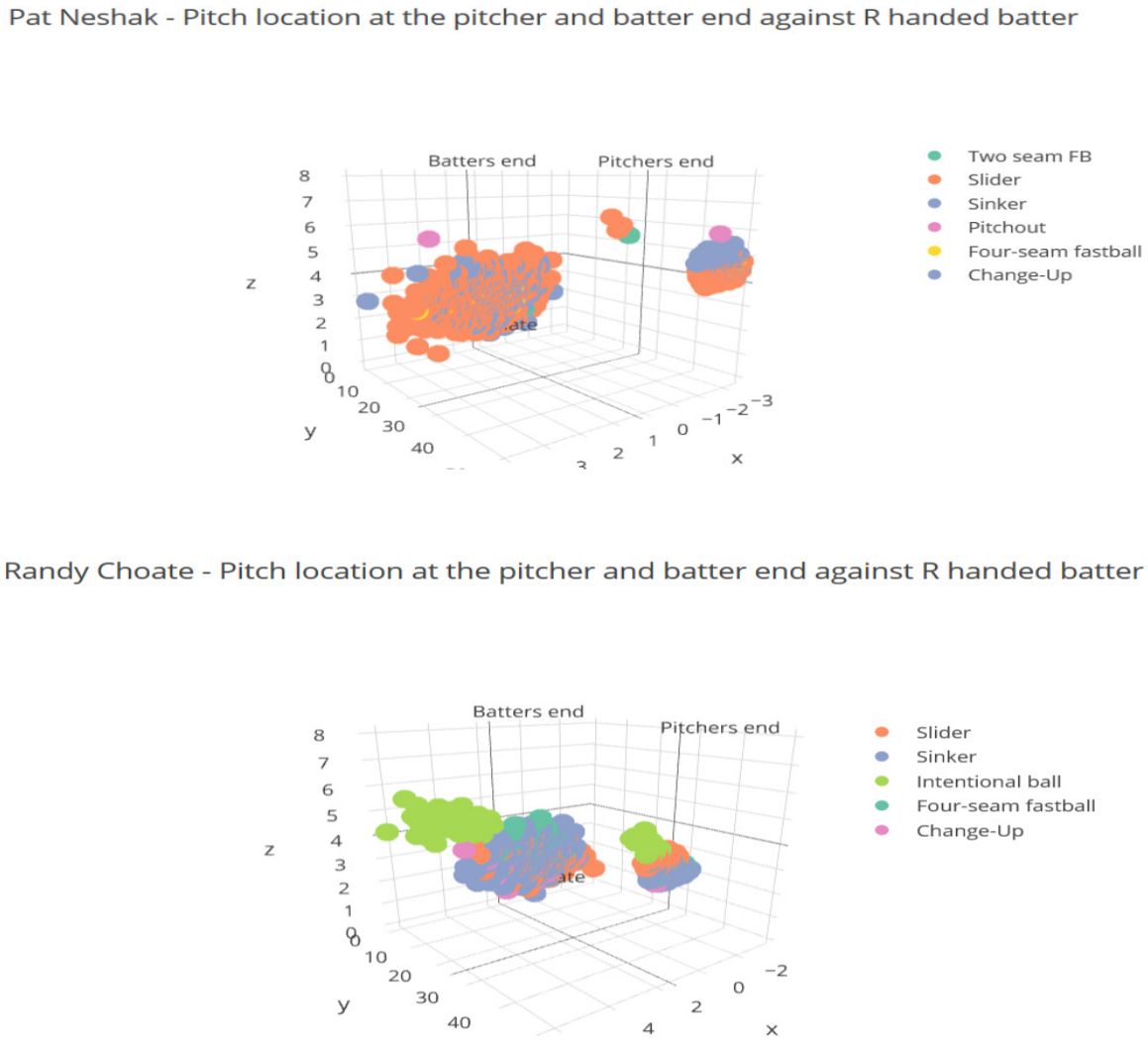
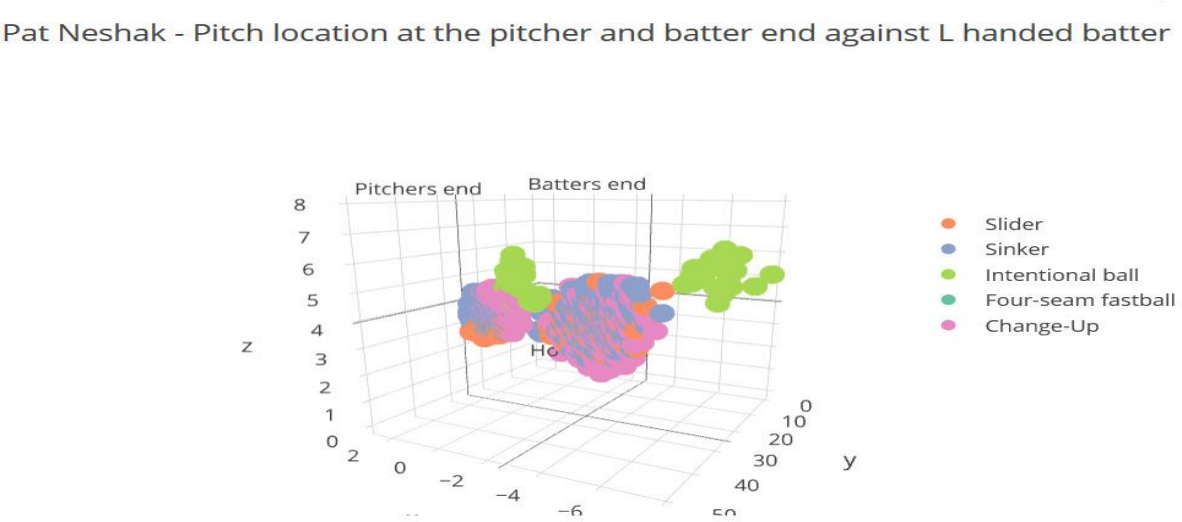


Fig 31. This figure shows the pitch location with various pitch types for Pat Neshak and Randy Choate against right handed batters.

To access the 3D feature of the image click - https://plot.ly/154/~anirudhkm/ - Pat Neshak

To access the 3D feature of the image click - https://plot.ly/150/~anirudhkm/ - Randy Choate.

In this 3D graph, the x axis corresponds to the right and left side of the home plate for positive and negative value. The y axis corresponds to the distance from the home plate origin. The z axis corresponds to the height of the pitch from the ground level.



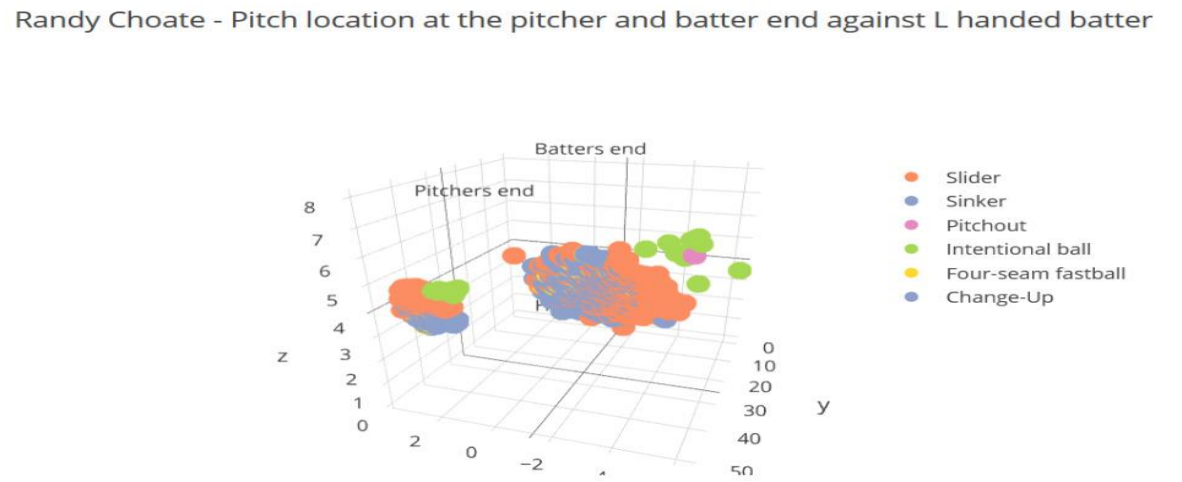


Fig 32. This figure shows the pitch location with various pitch types for Pat Neshak and Randy Choate against left handed batters.

To access the 3D feature of the image click - https://plot.ly/156/~anirudhkm/ - Pat Neshak

To access the 3D feature of the image click - https://plot.ly/152/~anirudhkm/ - Randy Choate.

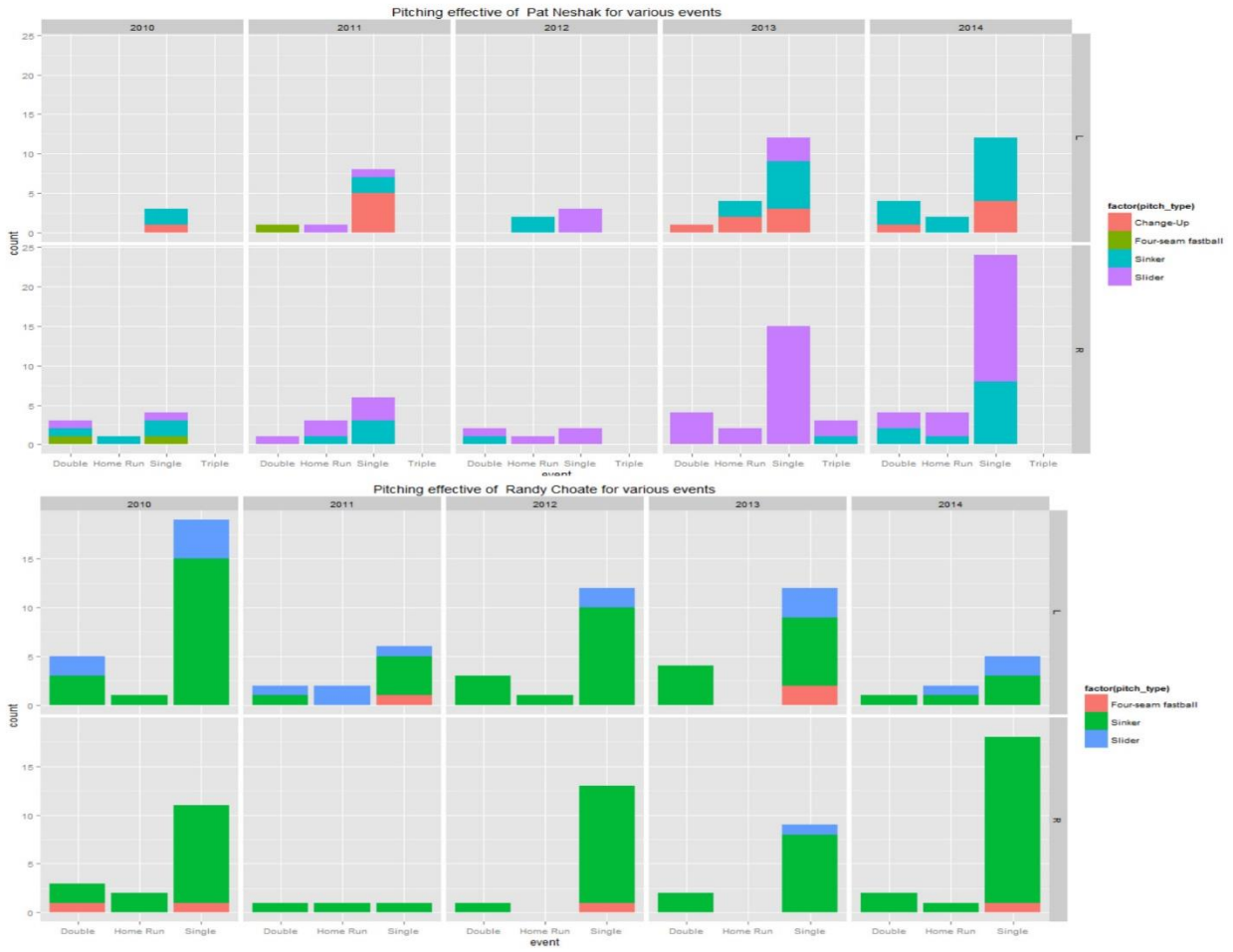


Fig 33. This figure shows the amount of single, double, triple and home run conceded by Pat Neshak and Randy Choate against right and left handed batters from the year 2010-2014.

From this figure we could see that Randy Choate has ever conceded a triple in the years 2010-2014. Also we can see that he has conceded most of the singles, doubles and home run with the sinker pitch type.

On the other hand we can see that Pat Neshak gave given singles, doubles, triples and home run with the pitch type slider mostly for the right handed batters and prefers change up and sinker deliveries to the right handed batters.

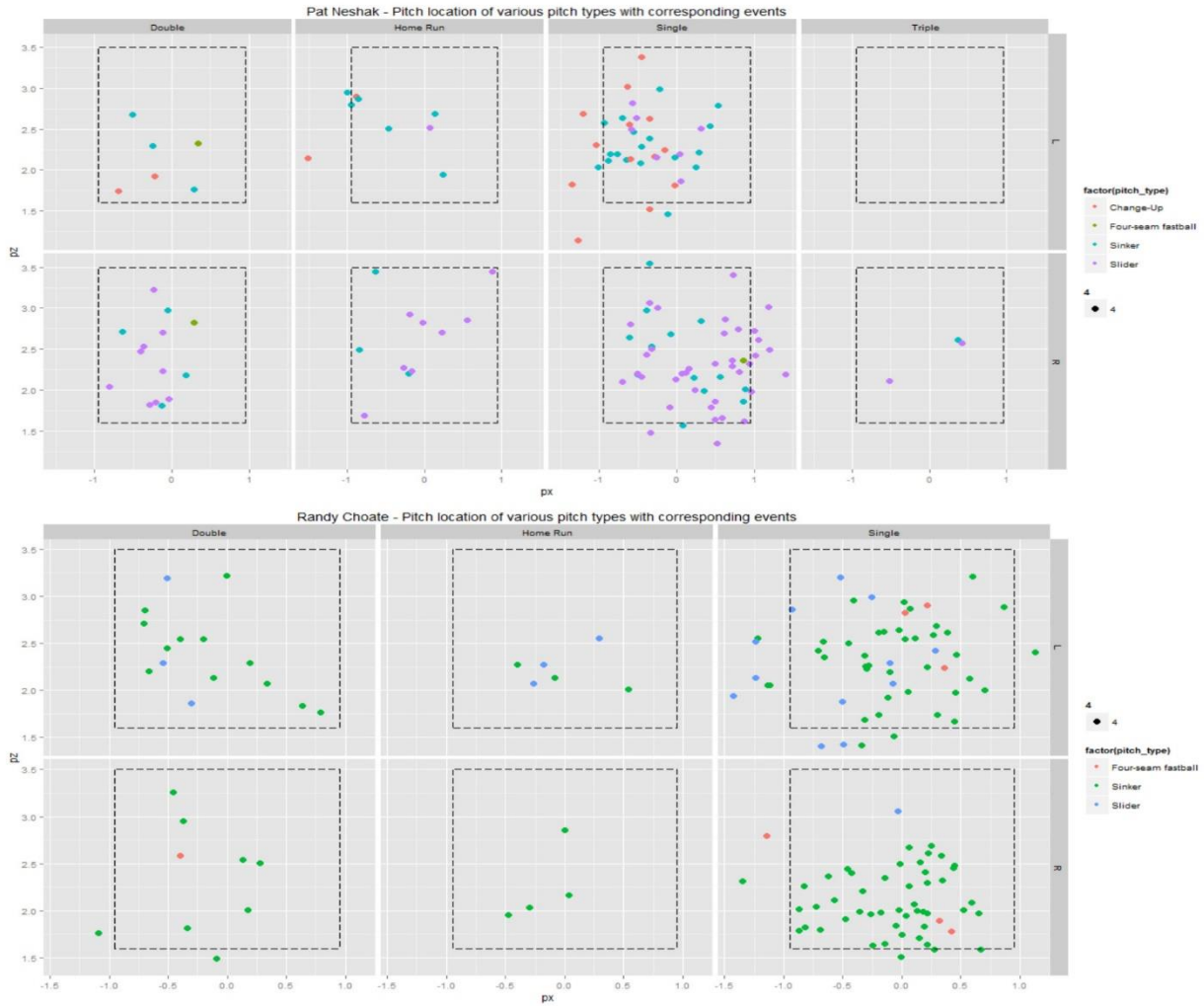


Fig 34. This figure shows the pitch location at the home plate for Pat Neshak and Randy Choate with the pitch types against right and left handed batters.

In this graph, a positive value for px corresponds to the pitch location on the right side of the home plate from the middle and pz corresponds to the height of pitch from ground level. A negative value of pz says that the pitch has bounced before it reaches the home plate.

Fig 35. This figure shows the number of flyout, forceout, groundout and lineout taken by Pat Neshak and Randy Choate against right and left handed batters from the year 2010-2014.

As we saw in the previous bar graph, we could see that Randy Choate prefers to pitch more sinkers and Pat Neshak prefers to pitch sliders more. As we saw before Pat prefers to mix between sinkers and change up for the left handed batters.

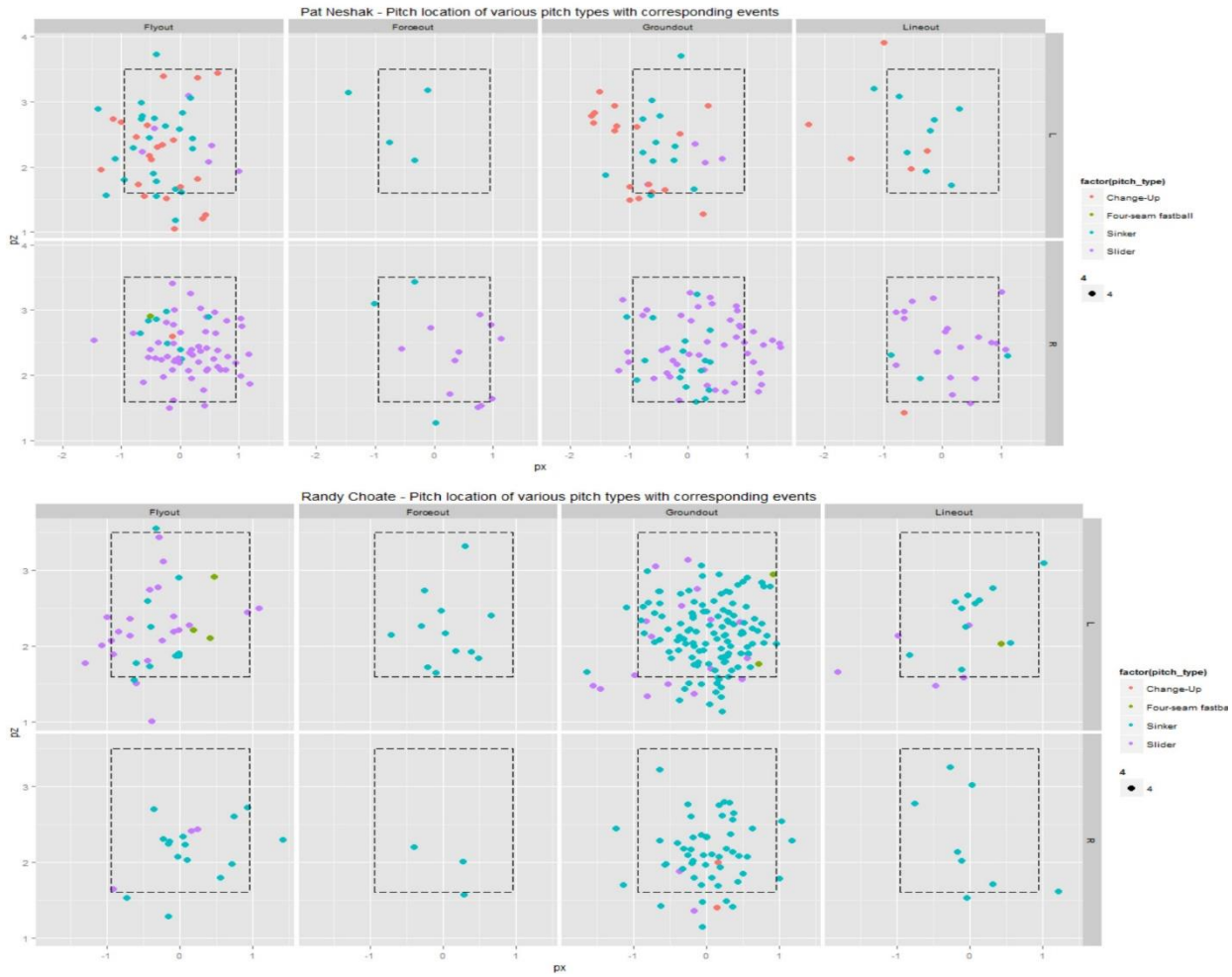


Fig 36. This figure shows the pitch location at the home plate for Pat Neshak and Randy Choate with the pitch types against right and left handed batters for the events flyout, forceout, groundout and lineout.

Pitchers analysis with Lahman's data

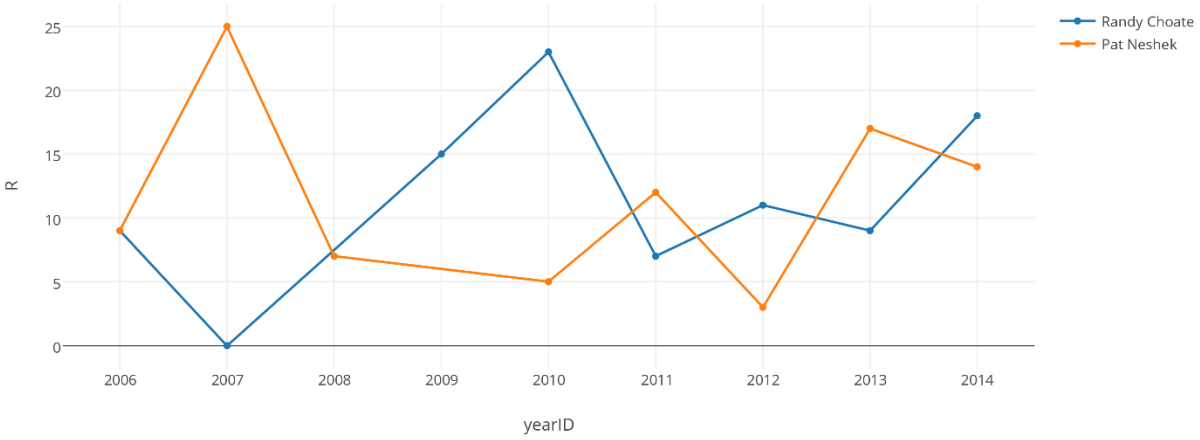


Fig 37. This figure shows the amount of runs conceded by Pat Neshak and Randy Choate throughout their career. Complete features of the figure can be accessed here - https://plot.ly/159/~anirudhkm/

From this graph we could see that Pat Neshak had given 25 runs in the year 2007 and Randy conceded only 0, this is because of Randy had played only two games in that season. After ups and downs in the middle period of their career, both of them shows similar type of results.

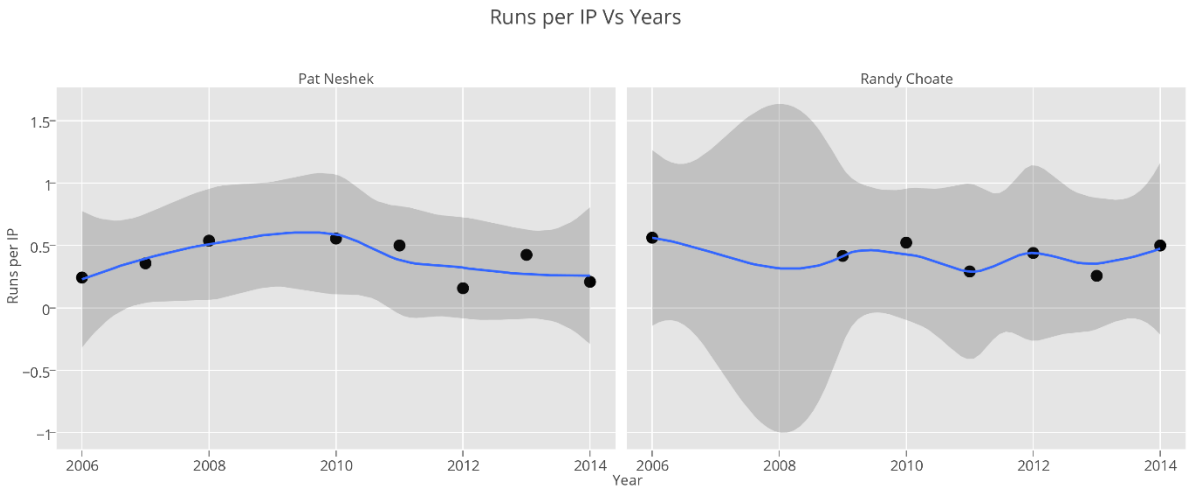


Fig 38. This figure shows the runs per innings for the pitchers Pat Neshak and Randy Choate throughout their career. Complete features of the figure can be accessed here - https://plot.ly/161/~anirudhkm/

We can infer that both the graphs shows a similar type of trend, we could see that the value hovers between 0.25 and 0.5. So we can say that both these pitchers show some good value when inferring the runs per innings value.

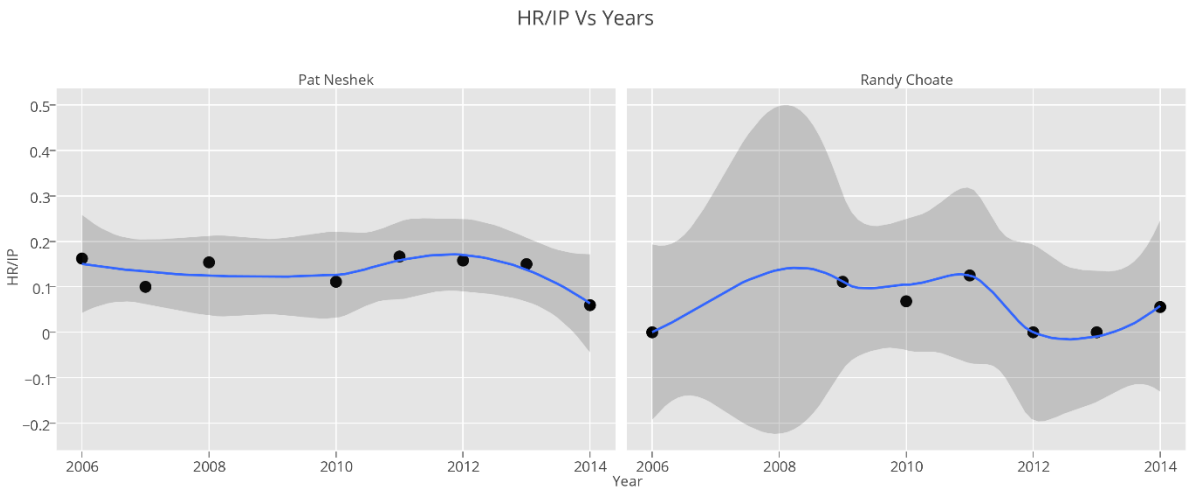


Fig 39. This figure shows the home run per innings value against the career years for Pat Neshak and Rany Choate. Complete features of the figure can be accessed here -https://plot.ly/163/~anirudhkm/

This graphs gives us an idea of which pitcher potentially concedes more home runs. We could see that Pat Neshak has almost a constant trend throughout his career with a decreasing trend from 2013.

In the Randy, we can say that most of the times the values are near zero and is always lesser than the value of Pat Neshak.

So from this graph we can say that Pat Neshak can concede more home runs compared to Randy Choate.

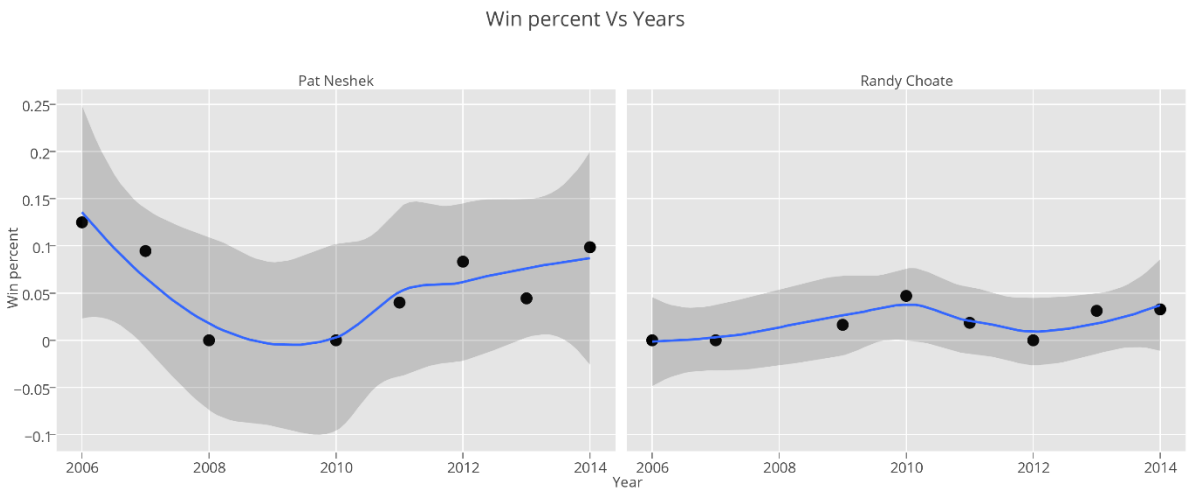


Fig 40. This figure shows the win percent for Pat Neshak and Randy Choate against various years. Complete features of the figure can be accessed here - https://plot.ly/165/~anirudhkm/

In terms of winning percentage we can see that Pat Neshak has a great decline in the initial career, but after the year 2010, we can see him gaining more winning percentage as year passes.

In the case of Randy Choate, we could see that his percentage is always less than 0.05 throughout his career.

So we could say that Pat Neshak is really improving his game after the year 2010, though he didn't have a great time during his initial career.

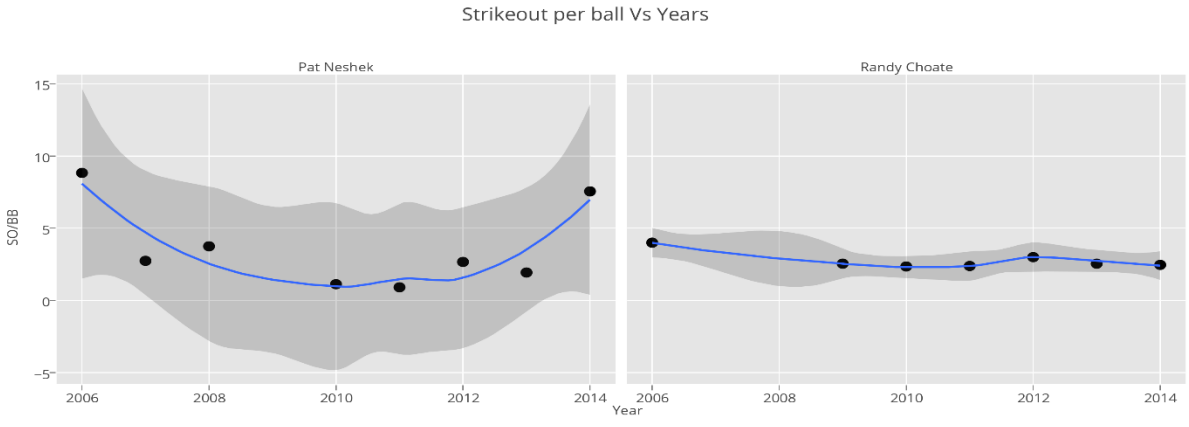


Fig 41. This figure shows the relation between strikeout per balls versus the years. Complete features of the figure can be accessed here - https://plot.ly/167/~anirudhkm/

From this graph we can see that Pat Neshak had a decrease trend which is quite bad for a pitcher, but after the year 2011, we can see that Pat Neshak has improved the value of SO/BB which is a positive sign.

On the other hand, we see that Randy Choate possess almost a linear trend throughout his career. But doesn't show any improvement like Pat Neshak.

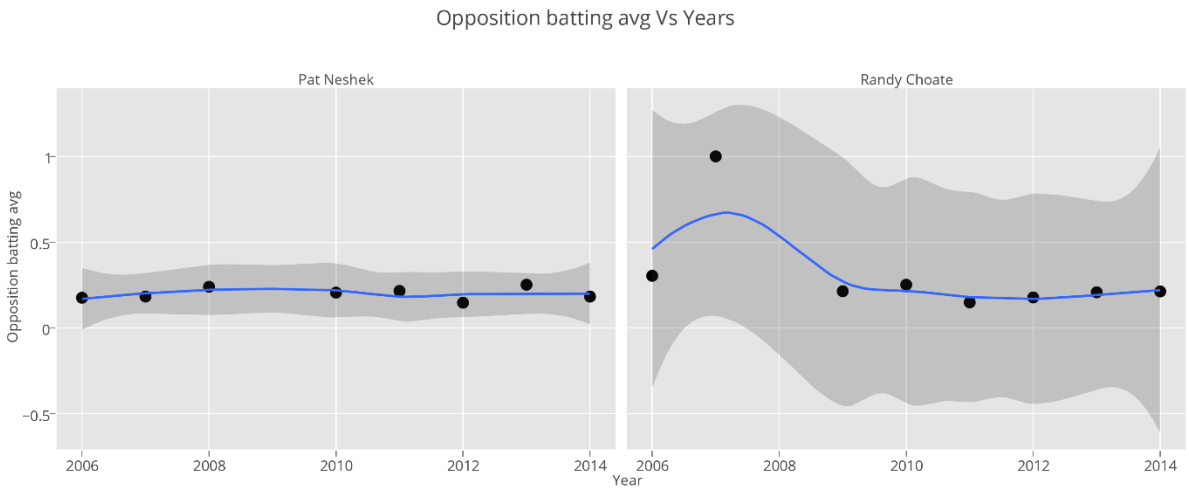


Fig 42. This figure shows the opposition batting average for Pat Neshak and Randy Choate against their career years. Complete features of the figure can be accessed here - https://plot.ly/171/~anirudhkm/

From this graph we can infer that, Pat Neshak has maintained a low opposition batting average which is very good for a pitcher, but Randy Choate has a higher opposition batting average and from the year 2009 has maintained a lower opposition batting average.

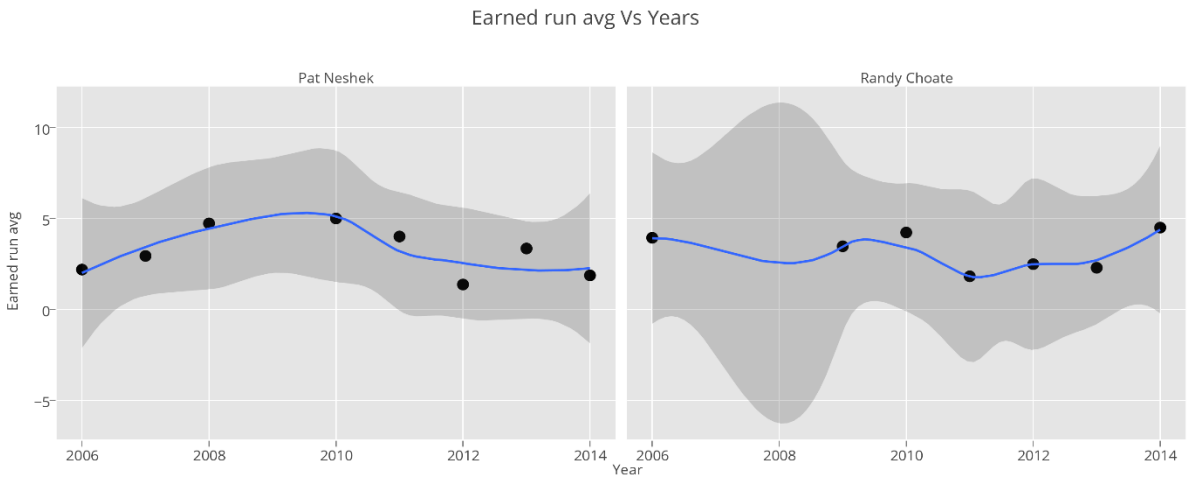


Fig 43. This figure shows the earned run average of Pat Neshak and Randy Choate against the various years. Complete features of the figure can be accessed here - https://plot.ly/173/~anirudhkm/

We can observe that Pat Neshak has an increasing trend in ERA initially and after the year 2010, we could see a decline trend which is really good for a pitcher. For Randy Choate, we can observe an inconsistent trend throughout his career.

Conclusion for pitchers analysis

From the above analysis we can conclude that Pat Neshak doesn't perform really well in the initial part of his career. But after the year 2010, he is showing improvement year by year.

But on the other hand, we see that Randy Choate performances is can be rated average throughout his career.

Final Conclusions and Future Directions

So these are few ways in which baseball data is being analyzed, so as to compare players and thereby help teams to win more matches.

Here are few more ways in which this project can be made more effective.

- These sabermetrics analysis can be applied to many players at a time, so as to compare their performances.
- These statistics can be applied to each game, so that we could have an insight about player's performance in more detail.
- These type of analysis can be done for other sports than baseball for the same purpose.
- Teams betting on other players can rely on these techniques, so that they will be confident that the new player will add value to the team.

References

- We thank *Mr. George*, who is a baseball player at Indiana University for giving valuable inputs about the game.
- Baseball statistics from Wikipedia, https://en.wikipedia.org/wiki/Baseball_statistics.
- Analyzing baseball data with R by *Max Marchi and Jim Albert*.
- Plotly for data visualization https://plot.ly/r/.
- ggplot2 for data visualization http://ggplot2.org/.
- A course on Sabermetrics - https://www.edx.org/course/sabermetrics-101-introduction-baseball-bux-sabr101x-0.
- Lahman's data set http://www.seanlahman.com/baseball-archive/statistics/.
- PITCHfx data set http://www.brooksbaseball.net/.