SportsMedia Technology Data Challenge

Cale Williams

Introduction

The purpose of this study is to assess infielder arm strength & accuracy using tracking data during routine plays.

Data Preparation & Reduction

Identifiers were assigned to each datapoint based on the features *game_str* and *play_id*. Only IDs that are in all of *ball_pos.csv*, *game_events.csv*, *game_info.csv*, and *player_pos.csv* (the intersection) were used in this analysis. This generated a quantity of 26,123 "plays", many of which are only pitches.

Then, plays with the following order of events were extracted:

1. pitch,
2. batted ball
3. caught ball by 1B, 2B, 3B, or SS,
4. thrown ball to 1B by 2B, 3B, or SS,
5. caught ball by 1B.

These are mostly ground balls to the second baseman, third baseman, or shortstop. They include double plays but the only throw being evaluated is the throw to first base. So if a first baseman catches a ground ball, throws to second base to get the lead runner, and the 2B/SS throws back to the first baseman, only the second throw is captured.

There are some outlier plays that include the desired events but should not be evaluated. Those will be inspected later. Additionally, errors and errant throws not caught by the first baseman are not included in the analysis.

Finally, 5 plays did not have ball tracking data available and were removed.

This filtering generates a dataset of 1,036 plays. This dataset serves as the baseline for all further analysis, unless otherwise specified.

Outlier Detection

As mentioned, some of these plays are not actually relevant. Recall that balls thrown to & caught by the *first baseman* were captured, not balls thrown to *first base*.

Below are locations of balls thrown to and caught by the first basemen. Note the few points from the shortstop and second baseman that are located in the outfield as well as locations in which the first baseman is nowhere near first base.

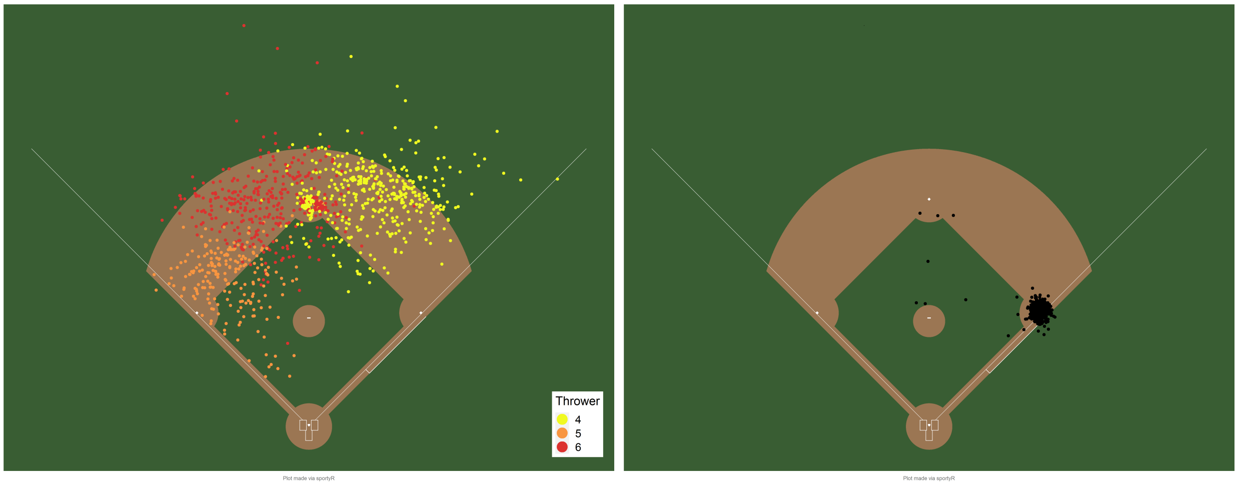


Figure 1: Where Balls to First Baseman are Thrown From (left) and Caught (right)

Here is one example of a play that may need to be removed, utilizing the ball tracking data to provide further context.

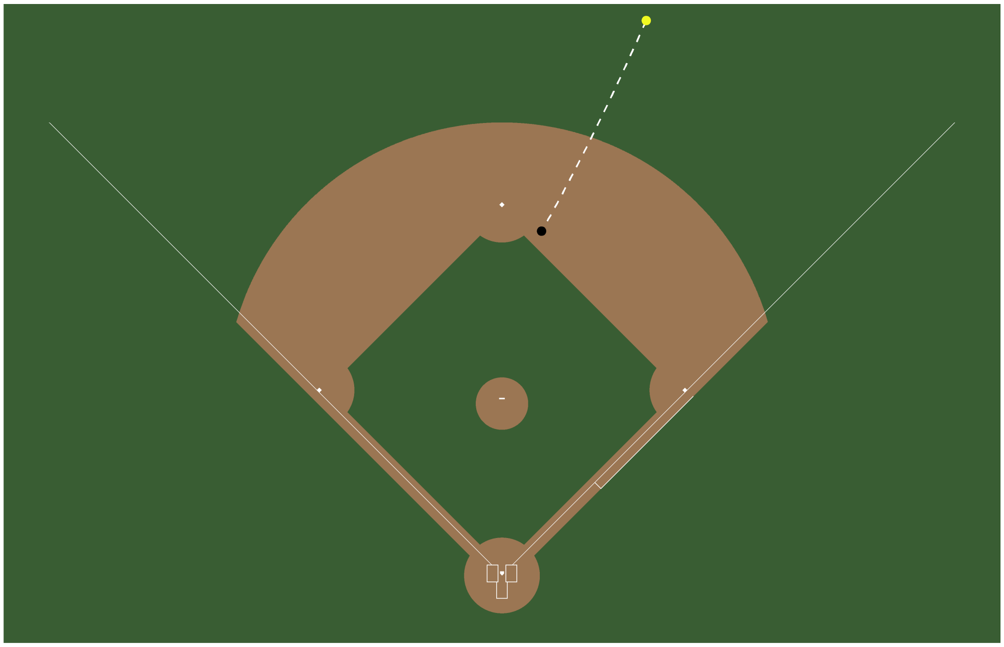
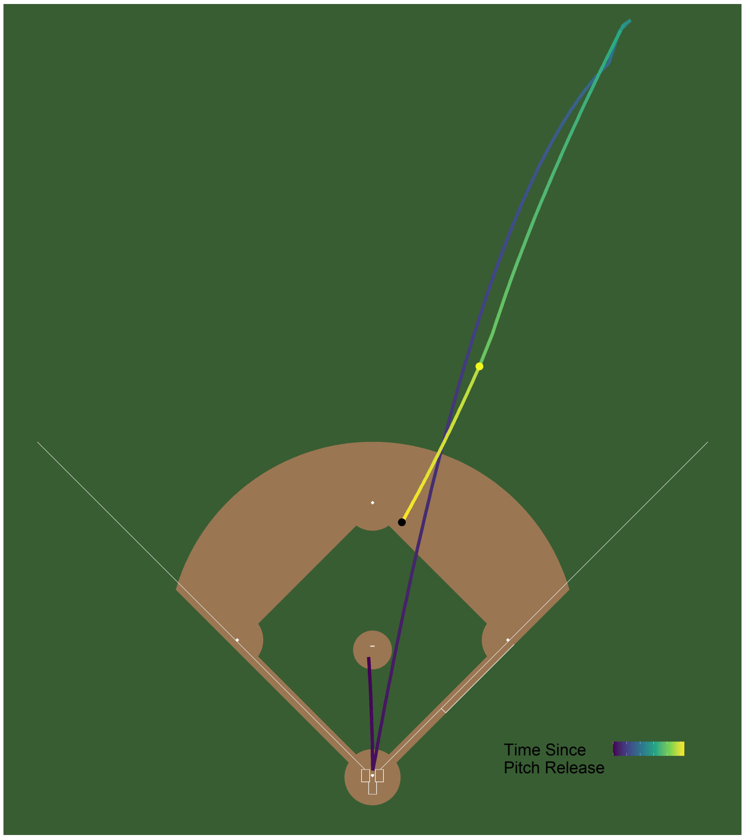


Figure 2: Potential Outlier

It is evident that this play satisfies the original constraints—the ball was caught by the second baseman and thrown to the first baseman. However, digging deeper reveals there's more that went on. Below is the ball location shown throughout the entire play, along with the logged events.



Time	Event	Player
0.0	Pitch	P
0.4	Batted Ball	Batter
4.7	Ball Bounce	-
5.6	Ball Bounce	-
6.3	Ball Off Wall	-
7.0	Ball Acquired	RF
8.0	Ball Thrown	RF
9.8	Ball Acquired	2B
11.6	Ball Thrown	2B
13.1	Ball Acquired	1B
13.3	End of Play	-

Figure 3: Play ID 1903_27_TeamNK_TeamB.186 Log

At the start of the play, there were runners on first base and second base. This appears to be a simple cutoff throw to the first baseman.

Similarly, here is an example of an errant throw by the shortstop. Note the sharp change in direction of the ball. The shortstop must have taken a few steps before throwing to first.

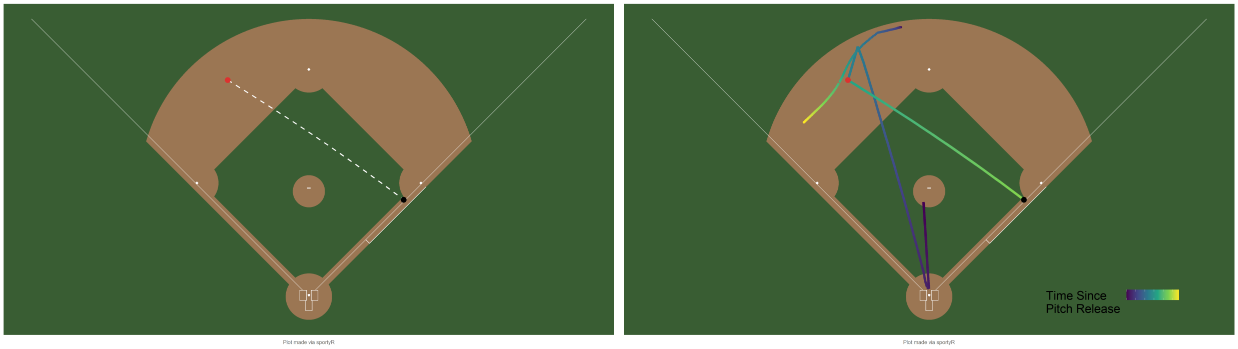


Figure 4: Play ID 1903_27_TeamNK_TeamB.186 Log

Because the goal of this data reduction is to identify routine throws to first base, these plays need to be classified differently and removed. Therefore any plays in which the first baseman catches the ball greater than 6 feet from first base are deleted. The dataset now consists of 914 plays.

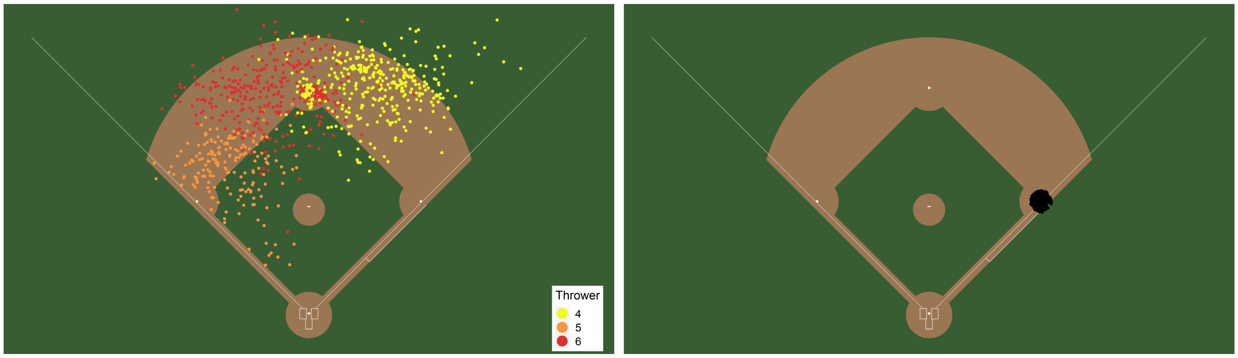


Figure 5: Where Balls to First Baseman are Thrown From (left) and Caught (right)

Throw Speed

The first aspect of arm quality assessed is speed of throws. Using ball tracking data, the average speed of a single throw, v, is calculated as:

$$v = \frac{d}{\Delta t}$$

where d is the Euclidean distance between the ball's location at the throw and at the catch and Δt is the change in time:

$$d = \sqrt{(x_{catch} - x_{throw})^2 + (y_{catch} - y_{throw})^2 + (z_{catch} - z_{throw})^2}$$

$$\Delta t = t_{catch} - t_{throw}$$

Below are the throw speeds by infielder.

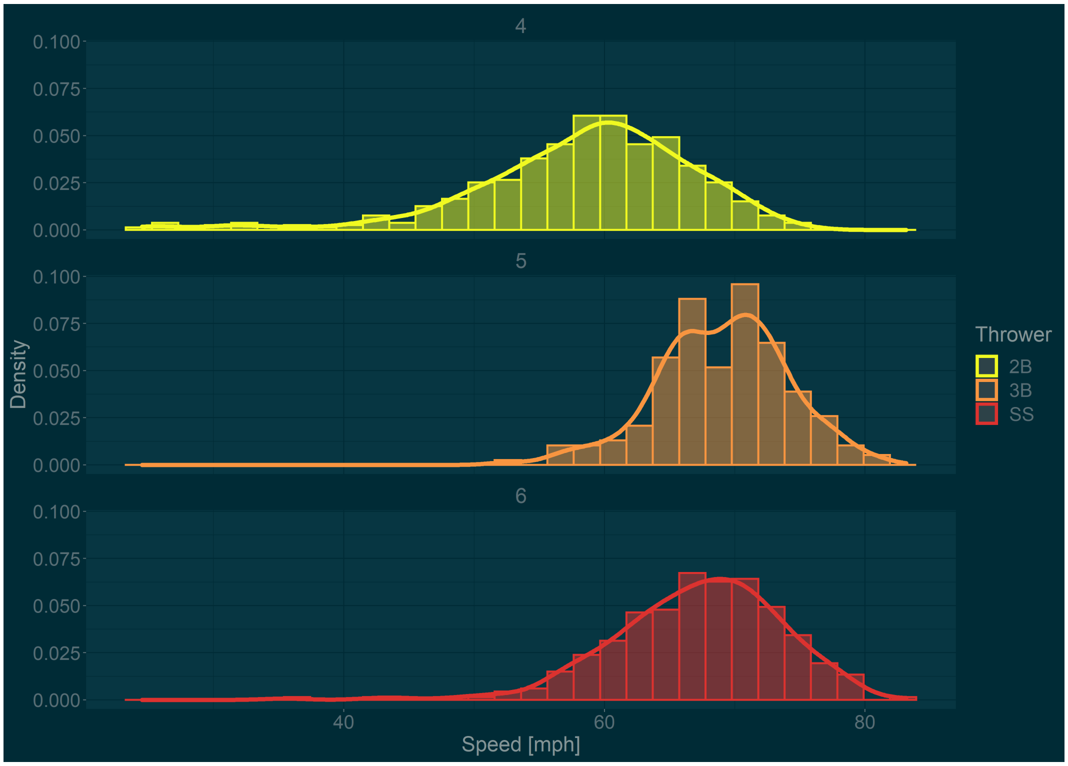


Figure 6: Speed of Infielder Throws to 1B

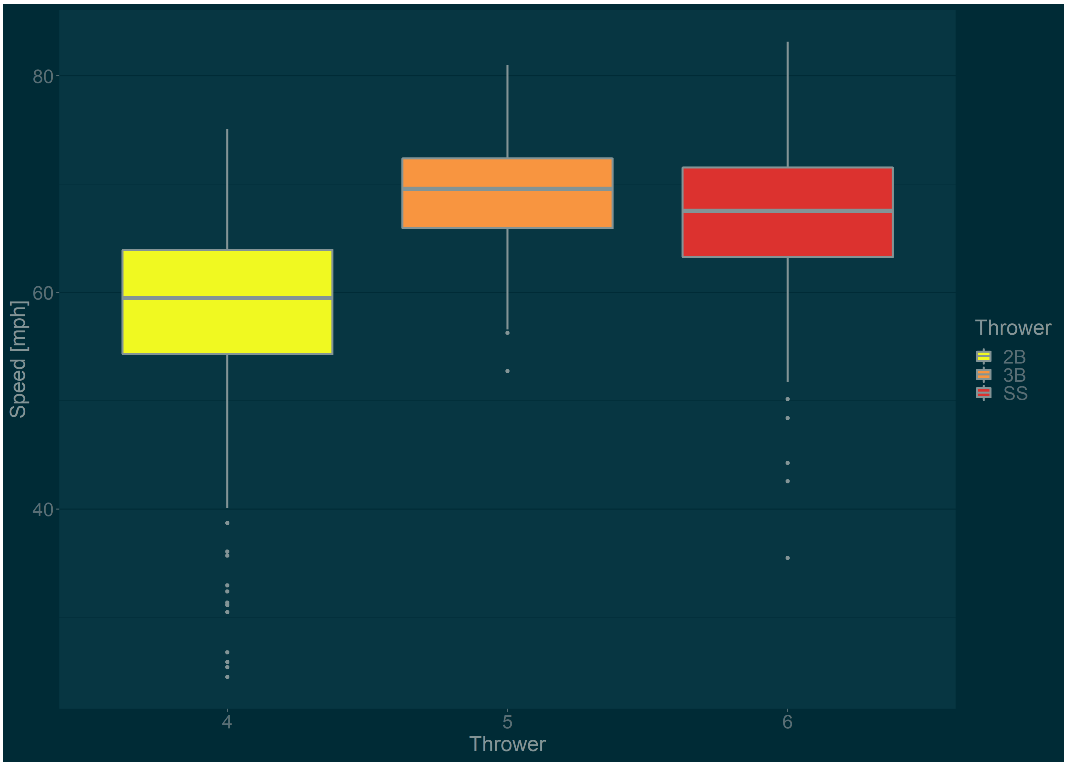


Figure 7: Speed of Infielder Throws to 1B

All distributions appear normal with negative skews. Third basemen throws also appears bimodal. Most notably, second basemen throw the ball with less zip than shortstops and third basemen.

Analysis of variance (ANOVA) can test this difference in speeds. An ANOVA model was built comparing the mean of speeds among infielder positions. The resulting p-value ≈ 0, therefore, the null hypothesis of equal means is rejected and the difference in throw speed between infielder positions is statistically significant. Additionally, F-value > 1 implying the differences between individual infielders within a position is larger than the difference between positions.

Is this difference a product of necessity or arm strength? As shown in the plots below, second basemen are usually closer to first base than the other infielders and thus do not need to throw the ball as hard to beat the runners. Or perhaps they play second base because they aren't able to throw it hard enough to play shortstop or third base.

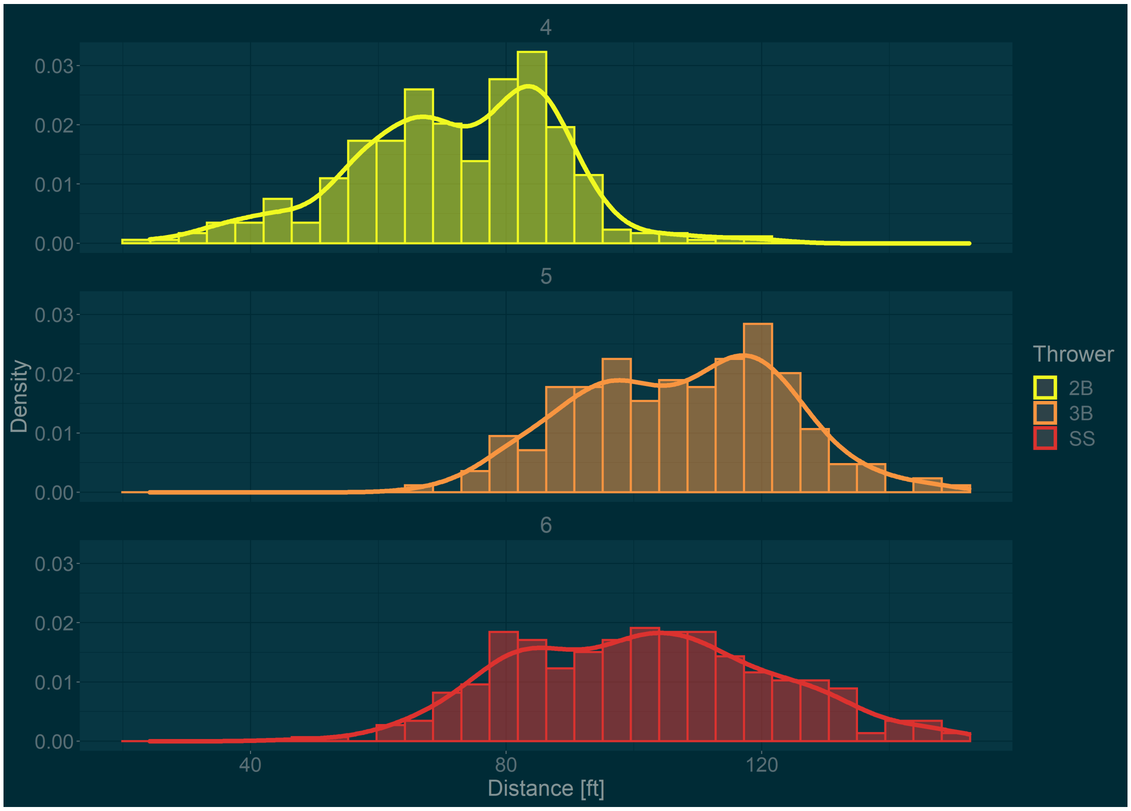


Figure 8: Distance of Infielder Throws to 1B

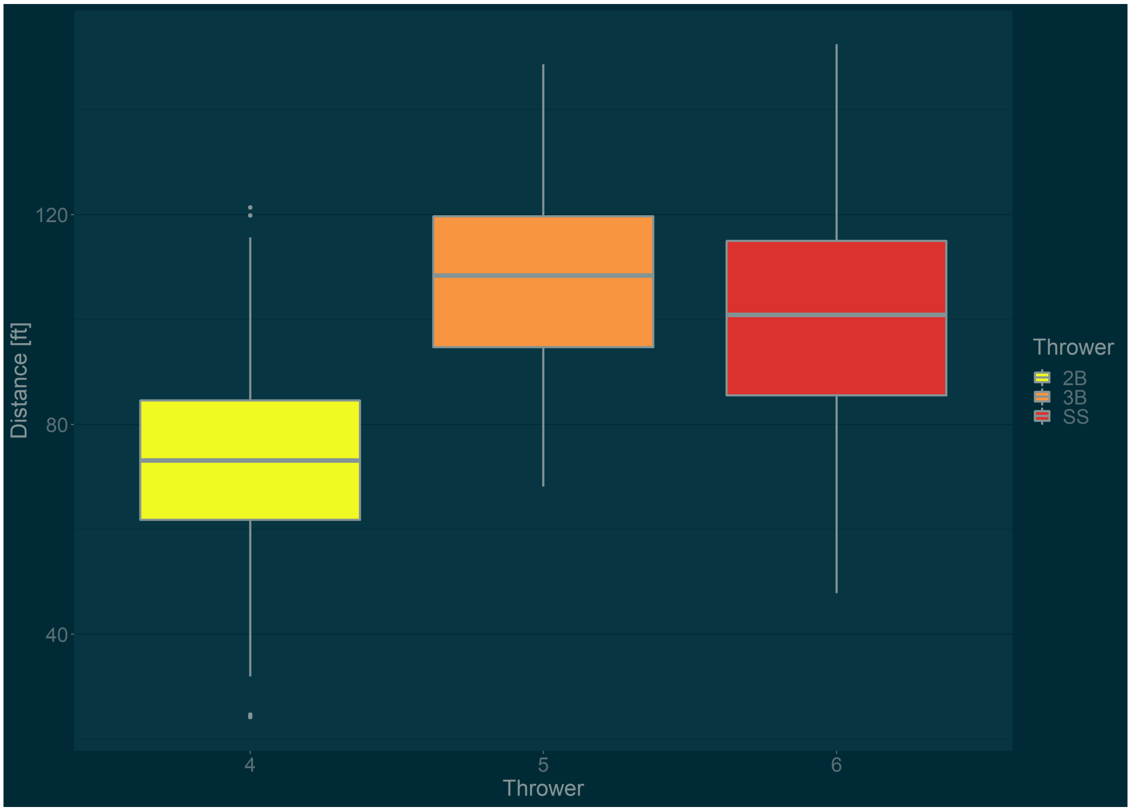


Figure 9: Distance of Infielder Throws to 1B

Next, “close” plays at first base were identified in an attempt to compare throw speeds when hard throws are necessary to get the out and remove plays in which a soft, easy throw is sufficient. Plays were classified as “close” if the runner got within 5 feet of first base within 1 second of the first baseman catching the ball. Admittedly, these values are fairly arbitrary, but can be easily changed if desired. 440 plays (48%) were classified as close. Shown below are the differences in throw speed between close plays and non-close plays.

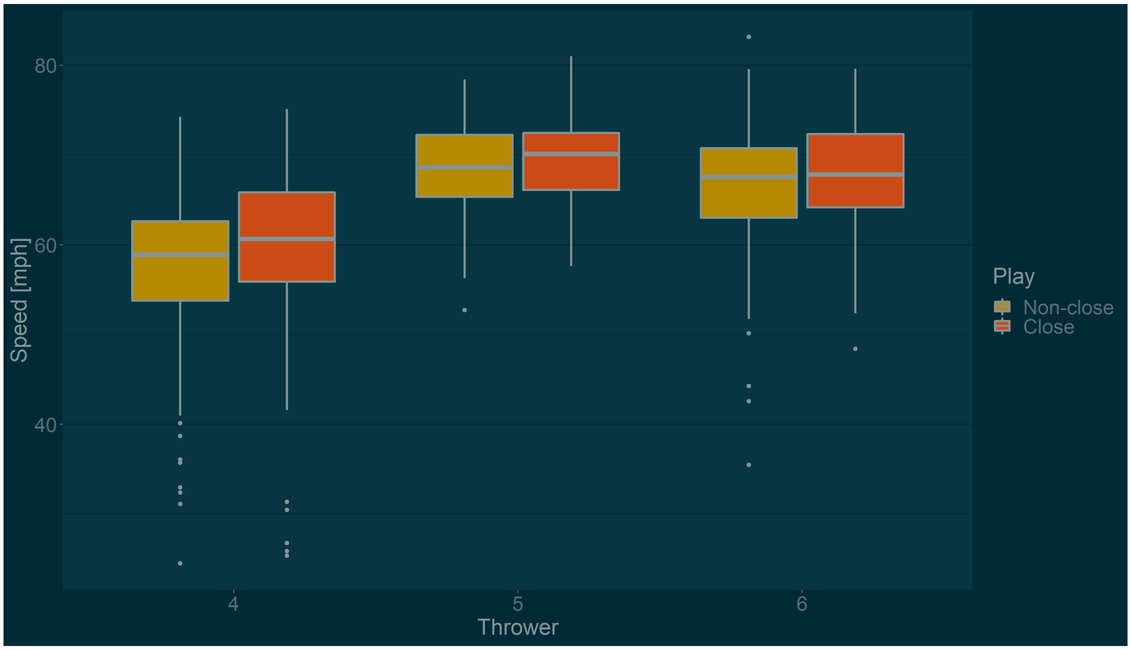


Figure 10: Throw Speed by Play Classification

The distributions are all shifted slightly up in speed, as expected. ANOVA between play types was performed for each position. The p-values for second basemen, third basemen, and shortstops are p = 0.02, p = 0.06, and p = 0.10, respectively. Thus, the difference in means is not statistically significant for third basemen or shortstops.

For the purposes of this study, this classification is deemed acceptable for second basemen and only close plays are used in the remaining analyses for second basemen only. Thus, it would not be meaningful to compare players among different positions, only those within a single position group.

The dataset now includes 700 throws to first base, with 178 from second basemen, 191 from third basemen, and 331 from shortstops.

Second Basemen

The players with the most throws to first base are shown below.

Player ID	Throws to 1B	Average Speed [mph]
2255	12	57
1628	8	59
9890	7	63
1201	6	57
2053	6	66
2235	6	61
2804	6	63
6189	6	65

Third Basemen

The players with the most throws to first base are shown below.

Player ID	Throws to 1B	Average Speed [mph]
1771	14	67
2382	9	70
2653	9	69
2818	9	69

Shortstops

The players with the most throws to first base are shown below.

Player ID	Throws to 1B	Average Speed [mph]
1181	18	65
1650	17	65
1972	16	68
1643	15	69
1959	15	67
2614	15	67

Throw Accuracy

To evaluate accuracy, ball tracking data was used to create scatter plots of caught balls by the first baseman from the perspective of first base. Below are the coordinate systems prudent to this exercise.

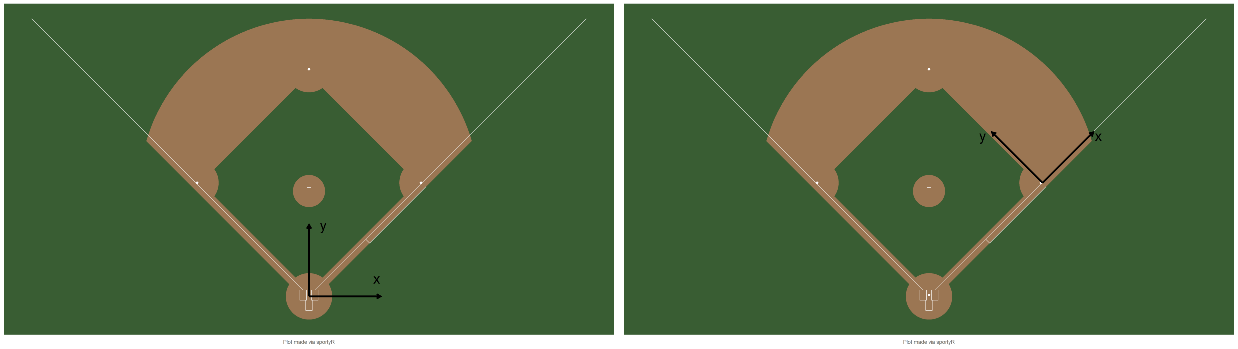


Figure 11: Raw Data CSys1 (left) and CSys2 (right)

The location of each thrown ball was extracted along with where it was caught, in *CSys1*, then rotated and translated to be in *CSys2* coordinates via the rotation matrix:

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

where θ is the angle between the two coordinate system's x-axes. This matrix rotates the x and y points about the z-axis. Below is a scatter plot of the results by thrower. The vertical lines represent the mean x catch locations.

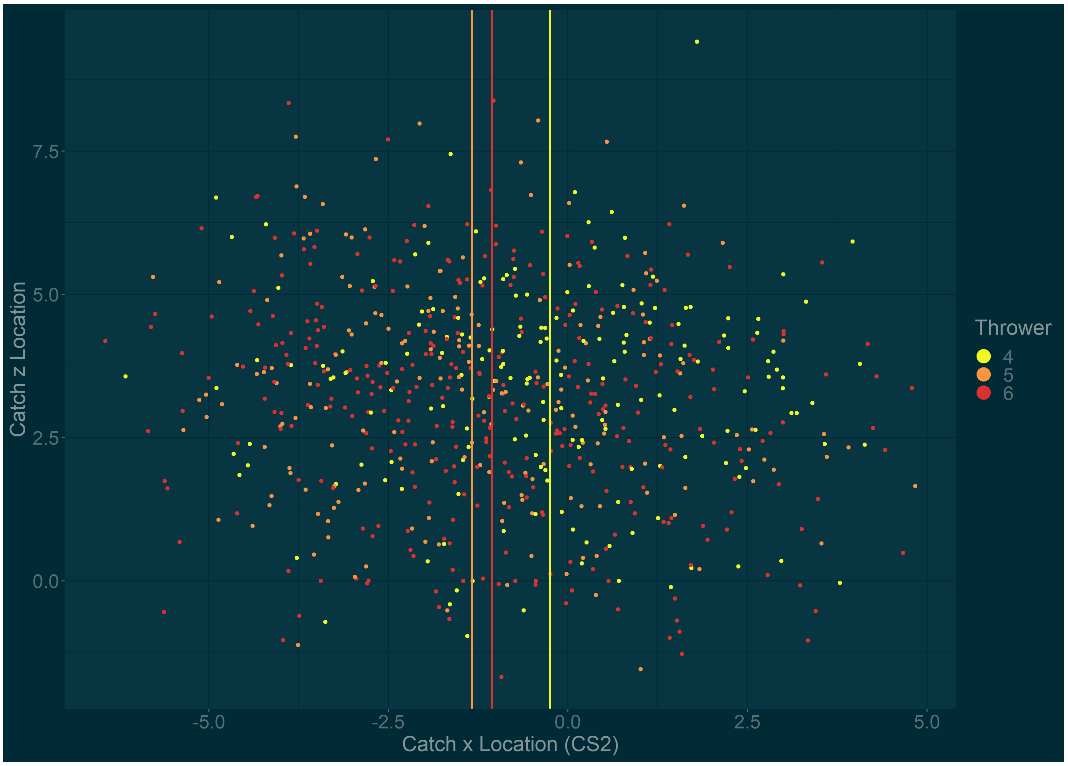


Figure 12: Where Balls are Caught, Relative to First Base

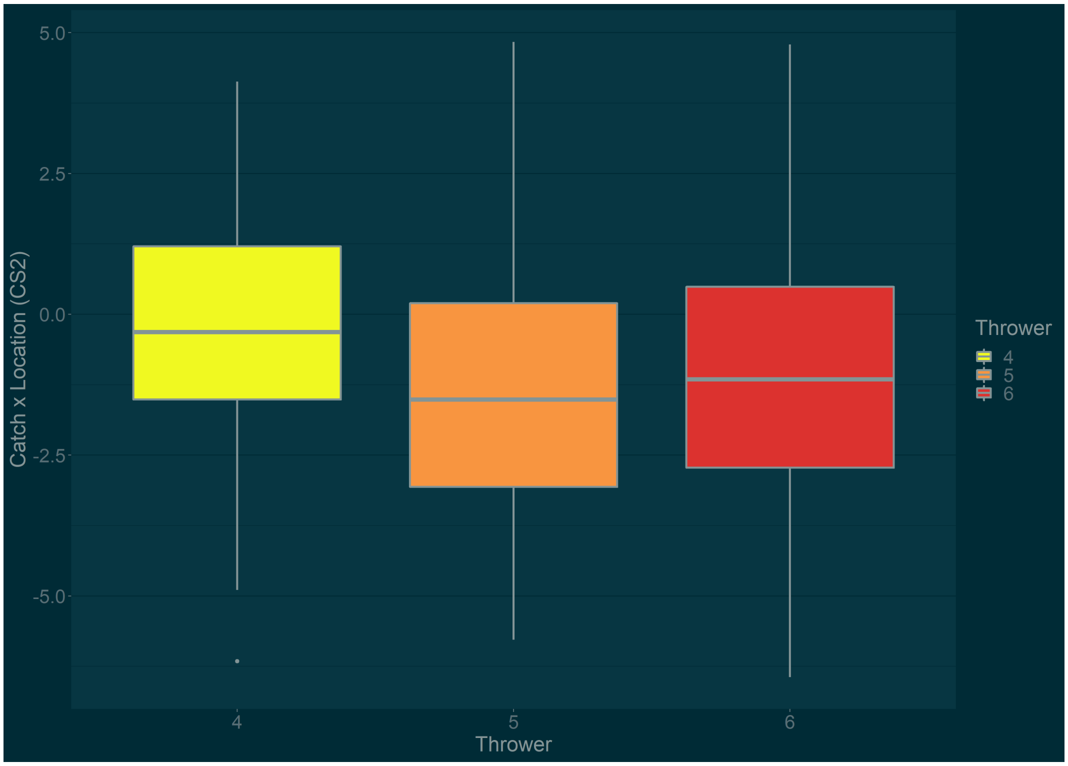


Figure 13: Where Balls are Caught, Relative to First Base

The differences are statistically significant. Note that $\mu_{3B} < \mu_{SS} < \mu_{2B}$. This makes sense as third basemen are more often throwing from the left side of first base. Therefore, to more correctly compare side-to-side throw accuracy between positions, this data needs to be rotated to body coordinates, or relative to the first baseman. To accomplish this, the angle was measured between the *CSys2* x-axis and the ball's location when it was caught. Assuming the first baseman always faces the direction the ball is coming from, the caught ball locations can be rotated via the same rotation matrix as before to be from the perspective of the first baseman, not of first base. This assumes the first baseman rotates around the z-axis in *CSys2* and are referred to as body coordinates. Below are the updated results.

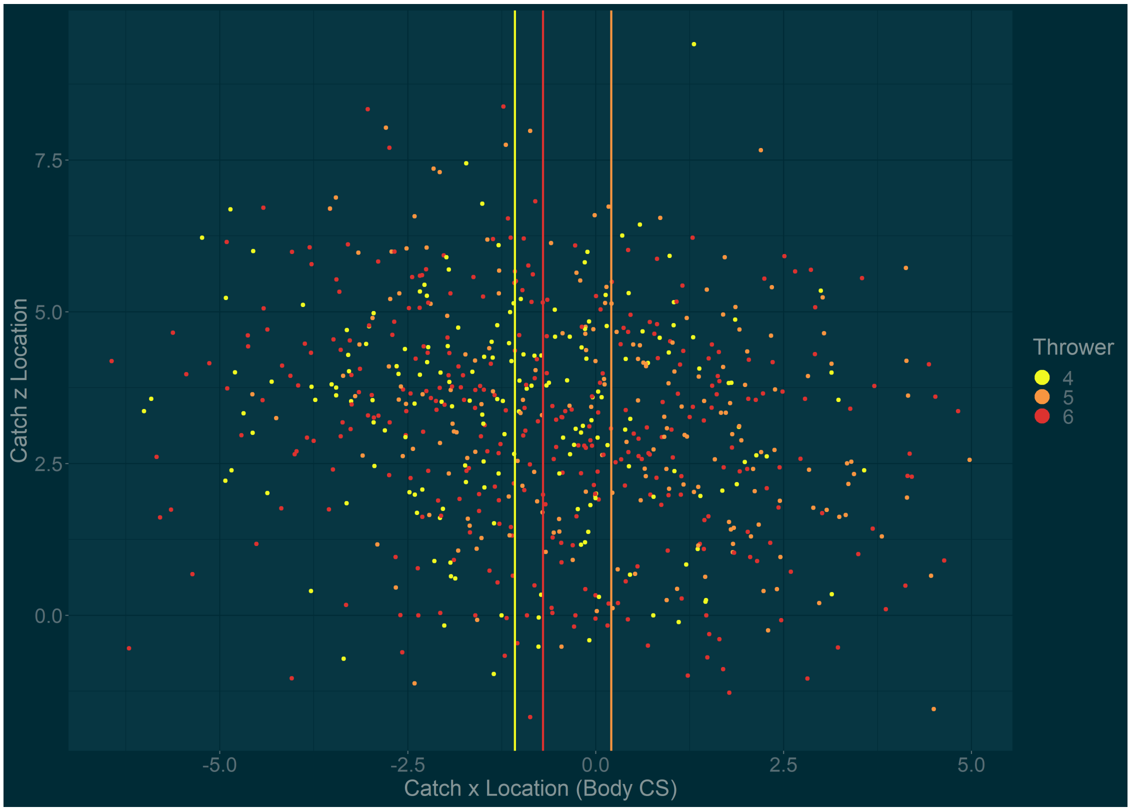


Figure 14: Where Balls are Caught, Relative to First Baseman

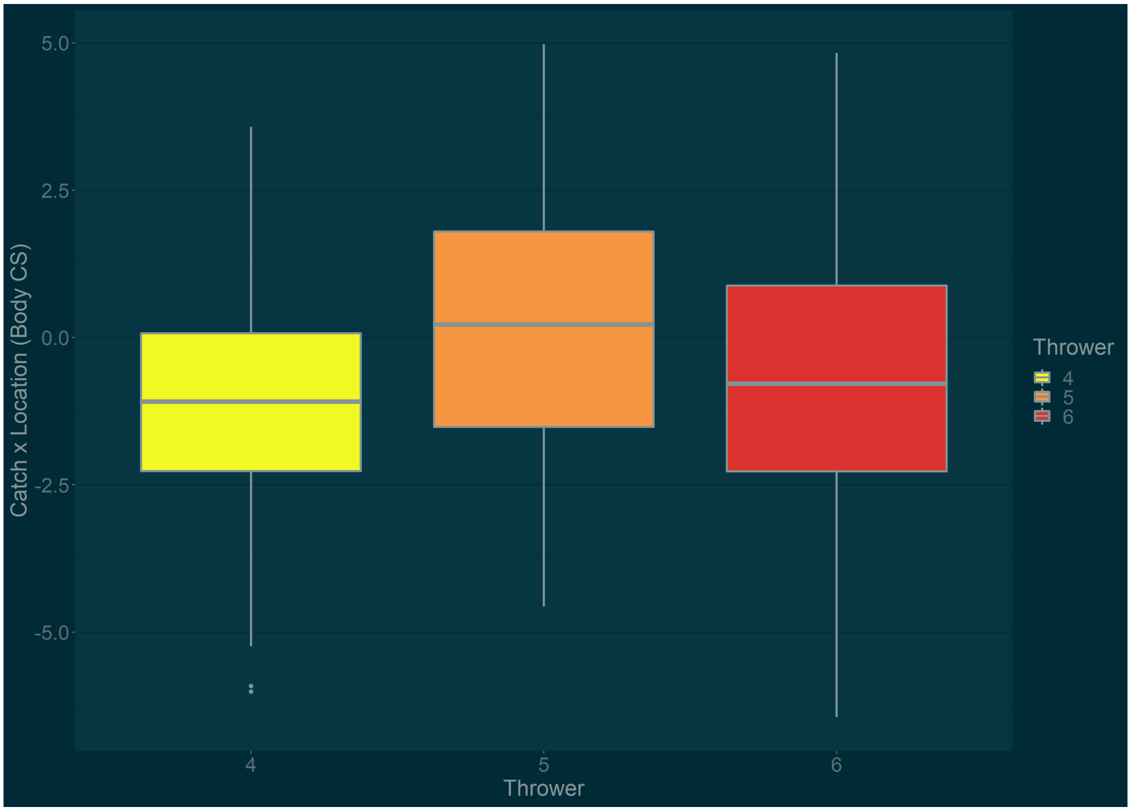


Figure 15: Where Balls are Caught, Relative to First Baseman

As expected, the catch locations have shifted closer to the $x = 0$ line for third basemen and shortstops. Additionally, the differences in means between positions are now not statistically significant as all ANOVA p $\approx$ 0. Given the assumptions, a throw at $x = 0$ implies it arrived in the center of the first baseman's chest whereas a throw at $x = -5$ is a throw that is to the left of the first baseman 5 feet and requires some significant stretch work by the first baseman.

These scatter plots have used the z-coordinates of the caught ball to represent height. However, because no origin is specified in the raw data for the z-axis, it is unclear what $z = 0$ corresponds to, and thus, is difficult to assess the z-accuracy of throws. This explains the negative z-values some caught balls show as well.

To assess accuracy, only the x-values in the body coordinate system are examined. The closer to $x = 0$ the throw is, the more accurate it is deemed.

Second Basemen

The players with the most throws to first base are shown below.

Player ID	Throws to 1B	Average First Baseman Catch x Location [ft]
2255	12	-1.1
1628	8	0.7
9890	7	-2.0
1201	6	-0.1
2053	6	-1.1
2235	6	-2.1
2804	6	-2.0

Player ID	Throws to 1B	Average First Baseman Catch x Location [ft]
6189	6	0.1

Third Basemen

The players with the most throws to first base are shown below.

Player ID	Throws to 1B	Average First Baseman Catch x Location [ft]
1771	14	-0.8
2382	9	-0.1
2653	9	1.5
2818	9	0.3

Shortstops

The players with the most throws to first base are shown below.

Player ID	Throws to 1B	Average First Baseman Catch x Location [ft]
1181	18	0.2
1650	17	-0.8
1972	16	-0.7
1643	15	-0.4
1959	15	-1.0
2614	15	0.6

Summary

The sample sizes are small for each infielder. Fortunately, many players appeared in games at more than one position. Below is a summary of the players with the most throws to a first baseman, regardless of position played.

Player ID	Throws to 1B	Average Throw Speed [mph]	Average First Baseman Catch x Location [ft]	Position(s)
1972	23	69	-0.4	3B, SS
1643	20	64	-0.5	2B, SS
1650	20	64	-0.9	2B, SS
1181	18	65	0.2	SS
1959	17	67	-1.0	3B, SS

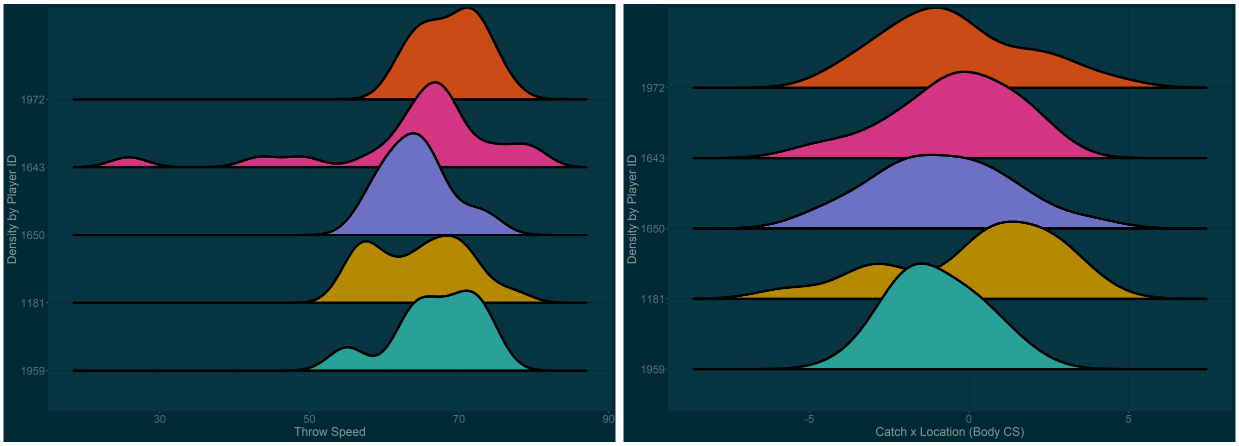


Figure 16: Players with Most Throws to 1B

Application

These methods could be used to assess both infielder arm strength & accuracy, especially with more data. Additionally, the ball y-location data in the body coordinate system could be added to the analysis to evaluate first baseman range.

Future Work

Close play classification: K-means clustering combined with principal component analysis was performed using features such as batter running speed, batted ball speed, and distance of throw to first. However, results were unsatisfactory with large variances within clusters and seemingly contradictory conclusions. Harder hit balls were more likely to be classified as close plays, for example. One would expect that a slow roller would allow the runner to get closer to first base than a hard hit ball. Perhaps for second basemen, only double plays should be classified as close plays. Another potential issue with identifying and filtering by close plays is that hard throwers may not be analyzed if their throws to first base are hard enough to not satisfy the current close play qualifiers.

z position: Standardization of z-locations would allow quantification of players' ability to throw balls at reasonable heights.