Report on Determining Home Field Advantage in Baseball

Isaac Gorelik, Nick Murphy, Pippa Hodgkins, Zoe Hsieh

May 2021

Problem Description:

Which stadiums in major league baseball produce the greatest home/away offensive and defensive splits by individual player ability, and if any show a significant difference in home vs. away performance, what causes that gap?

Initial Hypotheses

A. Home field advantage is an urban myth that may not actually exist.
B. It is not necessary that every team has the same 'level' of home-field advantage.
C. The cheating done by the Houston Astros in 2017 and Boston Red Sox in 2018 had a statistically significant impact on their batting behavior.

Data Collection

The beginning of advanced analytics in sports began with Major League Baseball, so luckily the data we need for this research is readily available from a variety of sources. We used Retrosheet's "game logs" and "play by play data" [1] for this research because logs from all of history are available in a complete, machine-readable form. We did not have to scrape their website for this data because all data is publicly available for download directly from their website. Therefore, any inaccuracies in our data are due to scorekeeping errors. We pulled data from the 2010 to the 2019 seasons for both the game logs and the play by play data.

Game Logs

Each Retrosheet log is a spreadsheet with 162 columns, with much information we do not need, including complete information and a day/night indicator, As we wanted to exclude any games which weren't completed, and the day/night indicator was redundant with the start time column, we began by eliminating these columns from our data. Other eliminated columns included the date and umpire information. Umpire information was

excluded due to the fact that umpires are not determined by the stadium, but instead, travel to games, so they are not an influencing feature of the stadium.

Play-By-Play Data

Each Retrosheet play-by-play file is another spreadsheet with 96 columns, with information on each play of each game of a season, for approximately 65,000 rows per team per season. We use this data to examine hitting patterns, so we can safely neglect irrelevant columns such as the batter's fielding position and other redundant fielder information.

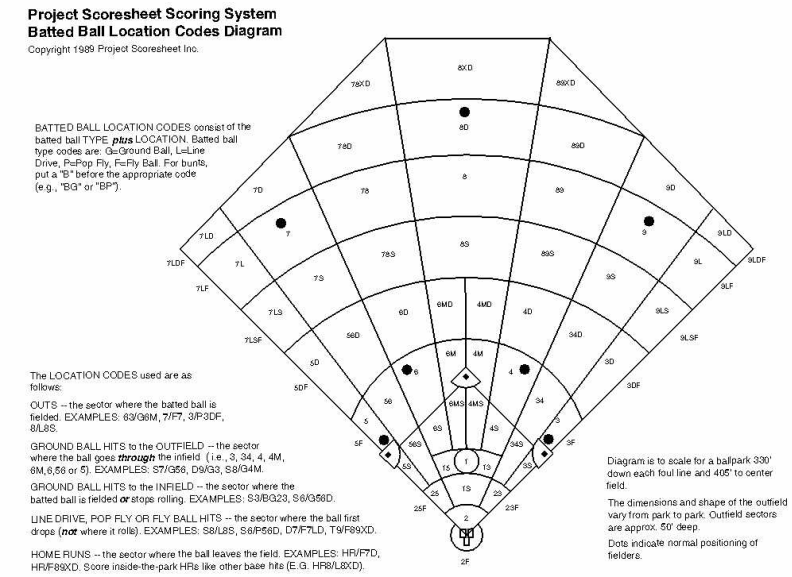
Data Processing

Game Logs

To process the game logs, we needed to be able to parse by inning, by the team, and by season. Therefore, we wrote functions to subset our data based on the desired attribute, allowing us to quickly perform high-level analysis. We are looking for long-term trends, so we combined our data for each team across 10 seasons.

Play-By-Play Data

To process the play-by-play data, we wrote a generic function that simplifies each play to a pair consisting of only the type of play it was and the play's hit location. The hit location is based on Retrosheet's Batted Ball Location Codes Diagram [1]:



Foul ball errors and Homeruns are assigned a hit location in the data, whereas batter interference (‘BINT’) and non-hits (‘K’, ‘W’, ‘C’, ‘HP’, ‘IW’, ‘SB2’, ‘SB3’, ‘K23’, ‘DGR’) were assigned to a hit location of 0. also inferred hit location from the type of play when needed, and made some simplifying assumptions. If someone was caught stealing (‘K’, ‘W’, ‘CS’, ‘PO’, ‘SB’, ‘DI’, ‘OA’, ‘PB’), this play’s hit location was interpreted as 0 (unknown). If a fielder made a basic play, we assumed that the hit location was the fielder’s location on the field. If the play was made by a fielder in foul territory, it was interpreted as a foul out. A failed bunt was assigned the hit location of the fielder who fielded the bunt.

Field Dimensions

Field dimensions for each team were gathered through Clem’s Baseball website [2]. Each stadium was listed by name and status, and we were able to get the estimated field areas for fair territory and foul territory, fence heights for left field, center field, and right field, the CF orientation, and Backstop. The site also had the outfield dimensions for left field, left-center, center field, right-center, and right field. We created a .csv file that contains each team’s stadium and their corresponding statistics.

Analysis

High-Level Offensive Analysis

PCA on Game Log Features

We started by running PCA on our dataset to determine if there was a low-dimensional representation of the data we could analyze, as our dataset initially comprised 96 columns worth of information, which we wanted to be able to condense. This presented some challenges, as our data contained both numerical and categorical (string-valued) data. At first, we addressed this issue by vectorizing our string-valued columns in the dataset. Given that many columns contained information with hundreds of categories (such as names of starting pitchers, names of players in the batting order, etc.), our transformed numerical matrix contained over 24,000 features but was extremely sparse. To address the difference in scale of all columns (number of fans in attendance, number of runs, etc.), we had to standardize each column. We tried two standardization techniques: scaling the data so it had mean 0 and variance 1, or scaling it so each point was between 0 and 1. The standardization method used did not have a visible effect on the projection of the data onto its first two principal components:

Figure 1: Scatterplot of Latent Representation of Count-Vectorized Game Data

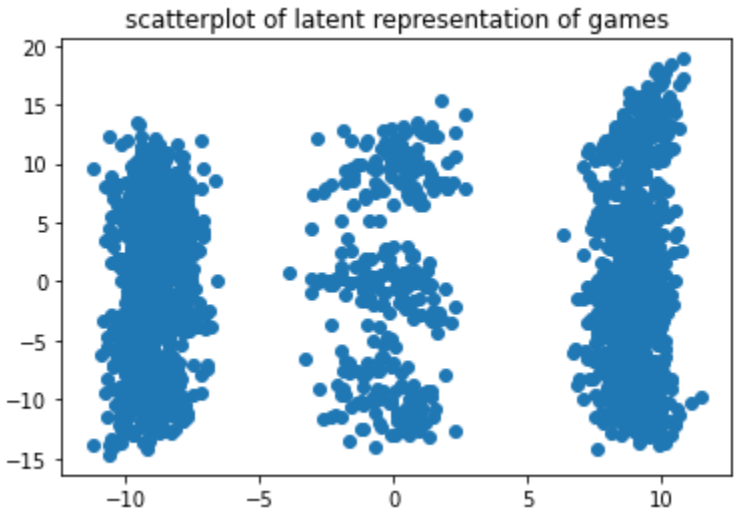


Figure 1 depicts the latent representation of all games from the 2019 season after running a principal component analysis. The two larger groupings are NL/NL games and AL/AL games, and the middle, sparser cluster represents interleague games.

Unfortunately, vectorizing our categorical features and then applying PCA did not yield any meaningful information–as shown, the four features which carried the most highly varying data were whether the home and away teams were AL or NL. As such, the

induced clusters did not give us any new insight not inherently obvious from the dataset. This unsatisfactory conclusion prompted us to abandon categorical features. The next step was running PCA using only the numerical features.

Figure 2: Scatterplot of Latent Representation of Numerical Game Data

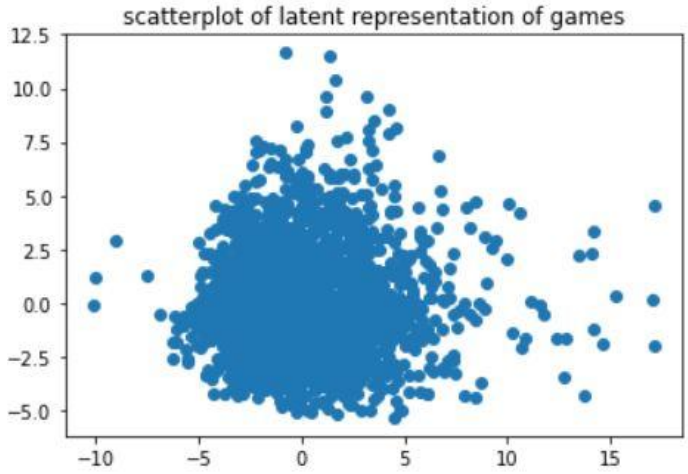
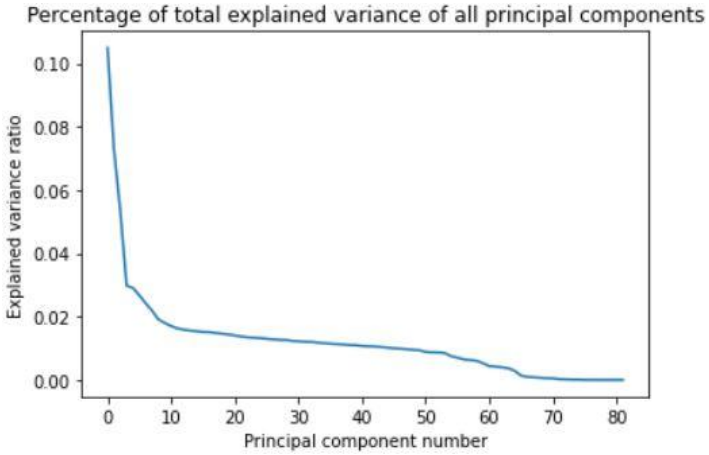


Figure 2 depicts the numerical data projected onto the two principal axes. Together, the first two principal components only account for roughly 17% of the explained variance–projecting onto two dimensions discards far too much of the high-dimensional structure. Given that there is no satisfying underlying structure to this "blob," we conclude that baseball game data has too much high-dimensional structure to be readily reduced to two easily interpretable axes.

Figure 3: Explained variance ratio of all principal components



The first principal component only accounts for 10 percent of the total variance. From this plot, we observe that we would need to keep roughly 50 principal components if we want our low dimensional projection to have 95% of the variance present in our original numerical game data.

As shown in Figures 2 and 3, the space of all baseball games was not particularly conducive to dimensionality reduction. We are working with intrinsically high-dimensional data, so we had to change our approach. We started by looking at the variation present in each offensive feature we had, starting with our runs scored. We

decided to look at the average runs scored at home vs the average runs scored away for each team, to see if there were any teams that were consistently scoring more at home than on the road, which might indicate some sort of advantage at their home field.

Average Run Differential:

Figure 4: Average Run Differential for Each Team

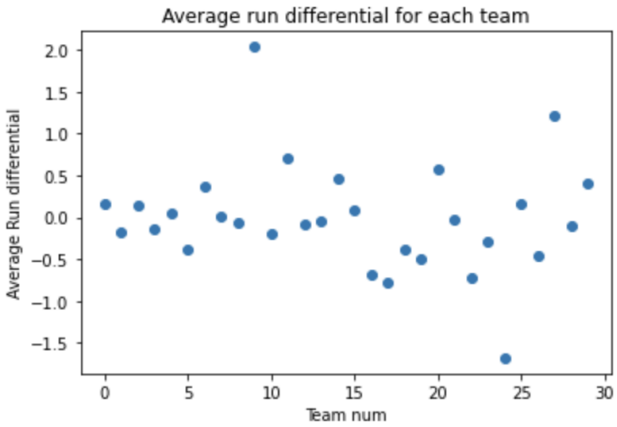


Figure 4 contains a scatterplot with every team in the majors on the x-axis (each team is assigned a team number in alphabetical order based on the three-letter team code assigned to them by major league baseball, so the Colorado Rockies (COL) are team 9) and the average run differential between each team's runs scored on the road and runs scored at their home field (average number of runs at home - average number of runs on the road) .

As seen in figure 4, there were several teams with a distinct difference in their offensive performance at home and away in both directions- the Colorado Rockies (team #9) had a clear increase in their offensive production at home, and the San Francisco Giants had a clear increase in their offensive production when on the road compared to being at their home stadium. In general, it appeared that most teams on average scored as many runs at home as on the road, though most teams fell on one side of the line or the other, as the likelihood of scoring the exact same number of runs home and away in one season is highly improbable. Next, we looked at the number of hits the home and away teams had during every game of the 2019 season to see if the average statistics were masking a game-by-game trend or if looking exclusively at the runs was being affected by teams getting unlucky with stringing hits together as opposed to genuinely low offensive production.

Unfortunately, our representation of the difference between the hits of the home team and the away team at any given game suffered from too much data noise. The sheer volume of points lead to an almost homogenous blob, with no real way to differentiate between the home and away team data points, or to determine a trend in either the home team or the away team's favor. We decided to change our approach slightly and switch back from

hits to runs (more directly related to the outcome of the game, which is what we care about) and instead of having each game have two data points associated with it in the graph, we have a single data point from the game which represented the run differential between the home team and the away team (runs scored by the home team - runs scored by the away team). By expanding our y-axis and halving the number of points visible on the graph, we hoped to present a more visually approachable way to determine if there existed a home team offensive advantage, resulting in Figure 5.

Figure 5: Scatterplot of Game Number vs the Difference in Runs of Home and Away Teams

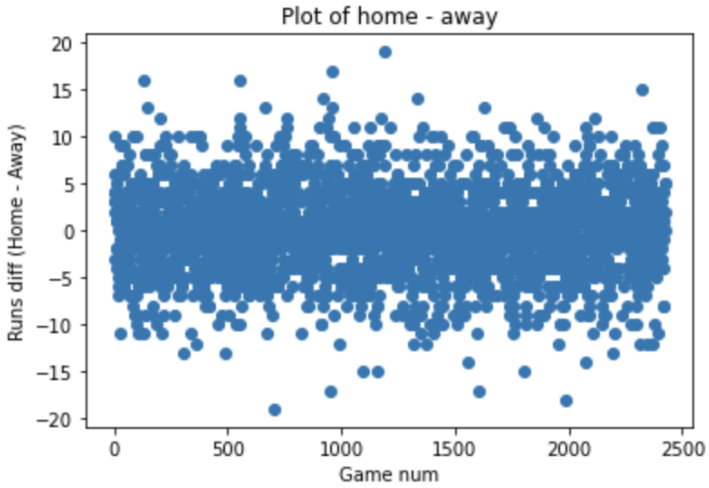


Figure 5 contains a scatterplot of every game in the 2019 season on the x-axis and the run differential between the home team and the away team at that game (home runs scored - away runs scored)

Figure 5 clearly shows that the run differential between the home and the away team is centered around 0; we can interpret this centering as neither the home team nor the away team generally scores more runs than the other at any given game. This centering of offensive production would indicate that home-field advantage as a general rule does not exist in major league baseball. Because the scatter plot has over 2,400 data points, we decided to reimagine the scatterplot as a histogram, with the bins representing the different run differentials to get a better sense of the curve, as due to the density of the points in the scatterplot it was possible that they were trending slightly in the home team direction in a way that was not visible in the blob.

Figure 6: Home vs Away Score Difference over Every MLB Game

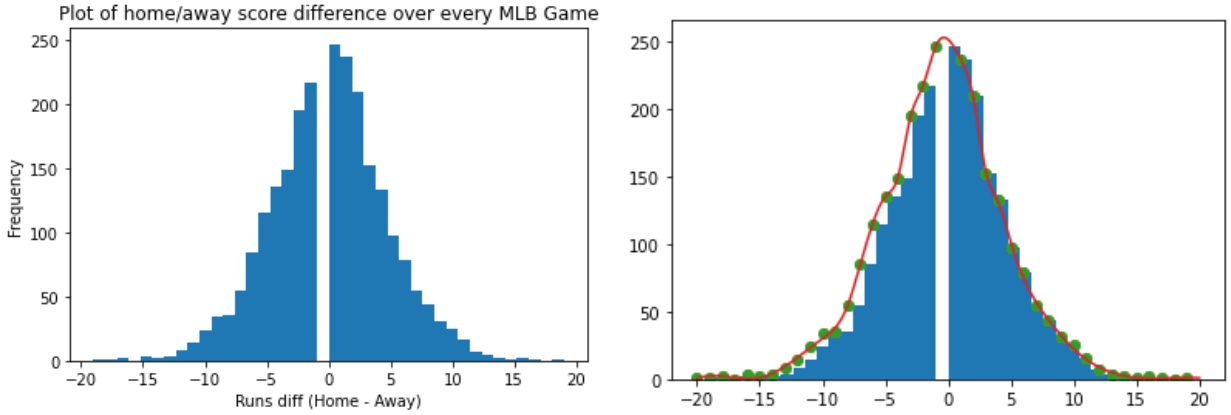


Figure 6 shows a histogram of the home/away run differential for every game in the 2019 season, with an interpolated curve overlaid based on the points taken from the histogram frequency. The histogram has a gap at a run differential of 0 because there are no ties in baseball, therefore there are no games where there is a 0 value for the run differential, leading to an empty bin.

After collating our run differential per game (home-away) for every game in the 2019 season with no differentiation by the team, we created an interpolated curve of the generated histogram to give a more visual interpretation of the general trend we found. To create the interpolated curve, it was necessary to exclude the bin density of the 0-bin, as including that datapoint would give any curve a steep drop at 0, whereas we wanted a more general trend line that didn't reflect the absence of ties in baseball. As seen in figure 6, the interpolated curve is centered roughly at 0, indicating that the run differential of home-away averages out to 0 across every game, meaning that the home teams in the 2019 season did not appear to have any sort of offensive advantage over the away teams in terms of runs scored across the entire league. This conclusion agreed with our previous observation in figure 6.

We then split this dataset into plots for each individual stadium, to test whether specific stadiums have differing responses to home and away teams. Upon splitting the dataset and calculating the home-away run differential per game per stadium, we observe in figure 7 that there are a couple of teams whose average run differential is either skewed left (the away team has a run production advantage) or skewed right (the home team has a run production advantage)

Figure 7: Interpolated curve of the run differential between the home team and the away team for every game at each home park in MLB

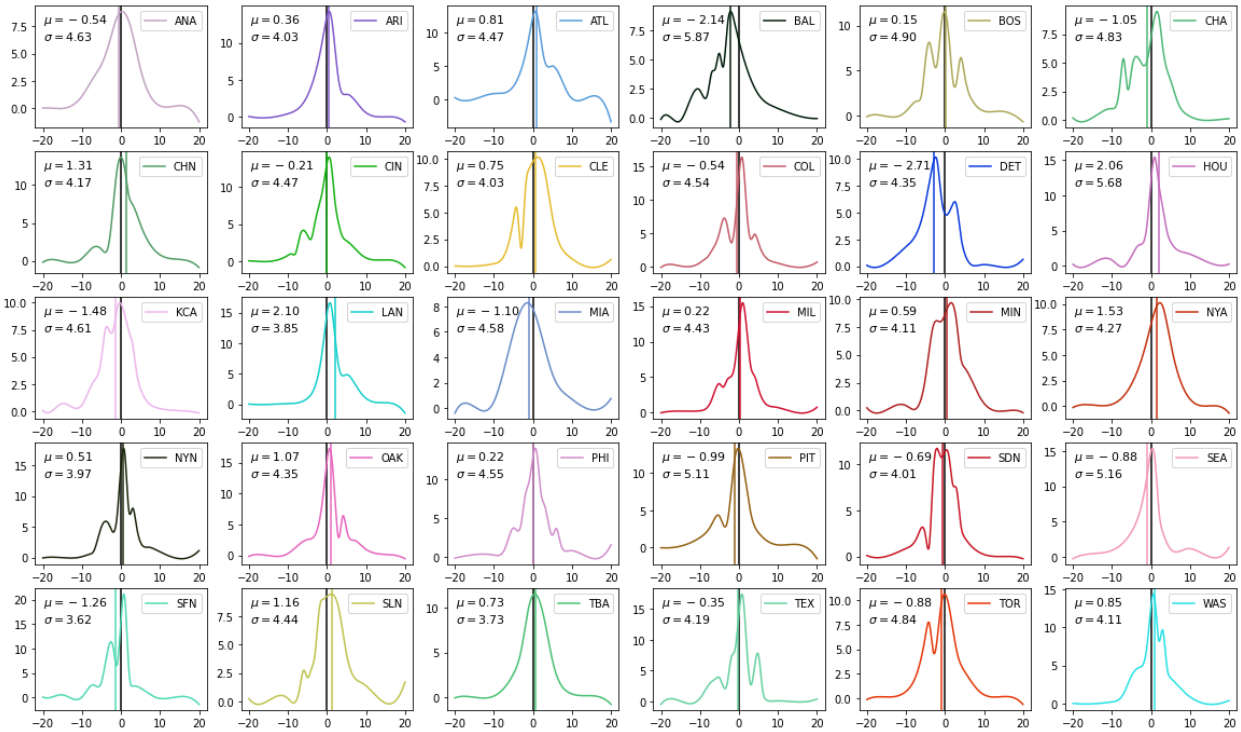


Figure 7 includes an interpolated curve of the histogram form seen in figure 6 of the run differential between the home and away team for each game played at the home stadium of each team. The figure also includes the mean run differential and a vertical axis at a differential of 0 to visually indicate which fields had a strong home or away run differential swing.

However, when we consider the dataset in the context of the 2019 season alone, the teams that have a strong left-leaning curve, the Detroit Tigers and the Baltimore Orioles, both had historically poor seasons in 2019, with both teams winning fewer than 35% of their total games- therefore the away team would be expected to score more runs than the home team at their stadiums due to the difference in quality between the teams, rather than due to a venue effect. The confounding factor of the quality of a team in a season proves a challenge when continuing with this project.

Hit Location Analysis

For the second half of this project, we decided to shift our focus from comparing the performance of the home team offensively to determining if there were any advantages/disadvantages to being gained from playing at a specific stadium. One of the most challenging aspects of analyzing any baseball offensive performance is that the stadium dimensions, specifically the amount of foul territory and the distances from home plate to the outfield walls, are not standardized. That variation in in-stadium playable territory is something we explored as the potential for small-scale advantages for

the home team, which would present themselves as an accumulated advantage over an entire season, as opposed to an advantage on a game-by-game season.

Foul Territory and Foul Outs

We began by analyzing if there is a relationship between the amount of foul territory in a stadium and the outcomes of plays where the ball is hit into foul territory. In order to accomplish this, we obtained and pre-processed the play-by-play data and the field dimension data. To record foul balls from the play-by-play data, we parsed the given fielding sequence, which was Retrosheet's method of encoding how the ball was caught and thrown during a play. A foul ball occurred when an 'F' was found in the pitching sequence. With this information, we appended a new column to our data that represents the number of foul balls per at-bat. A foul out occurred when the pitching sequence ended in an 'F', which means that the at-bat completely ended on a foul ball, which is only possible for a foul out. First, we plotted the number of foul balls by the stadium in order to check if any team was an outlier.

Figure 8: Scatterplot of Total Number of Foul Balls by Stadium in the 2019 Season

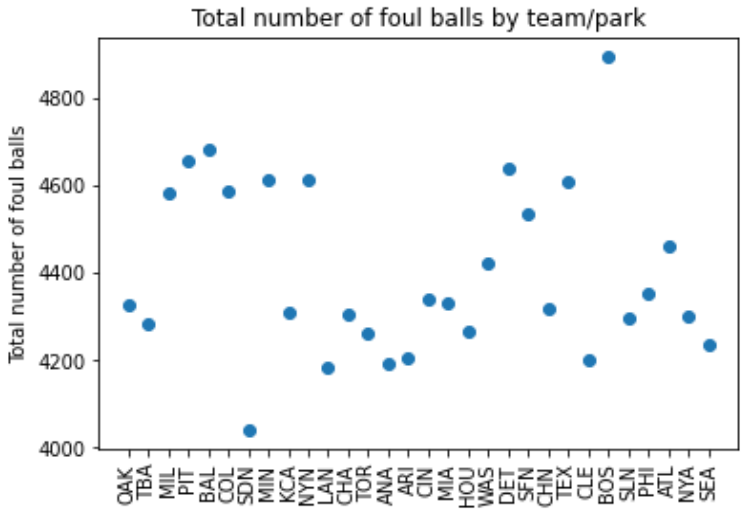


Figure 9 includes a scatter plot representing the total number of foul balls hit at each team's home stadium.

The scatterplot above shows two outliers: Fenway Park (home of the Boston Red Sox), which had over 200 more balls hit in foul territory than the next highest stadium, and Petco Park (home of the San Diego Padres), which had approximately 200 fewer balls hit in foul territory than the next lowest stadium. While we expected Fenway Park to have an increased frequency of foul balls as it has some of the smallest playable foul territories in the major league, leading to more balls hit foul to land in the stands, we did not expect Petco Park to have such a small number of foul balls over ten seasons, as it has comparatively an average amount of playable foul territory. However, our analysis of the total number of foul balls is skewed by the offensive tendencies of the home teams that

play most frequently at these stadiums: a stadium that is home to a team that makes more contact would have an increased number of foul balls due to the higher volume of contact. Additionally, an increased number of foul balls is not necessarily an offensive or defensive advantage/disadvantage for any team, but more foul-outs would be an advantage for the defensive team. We decided to analyze the total number of foul outs and the ratio of foul outs to foul balls- essentially determining if there were any teams that were turning more foul balls into outs, which would constitute a defensive advantage.

Figure 9: Scatterplots of Foul Outs and Foul Out Ration by Amount of Foul Ball Territory for the 2019 Season

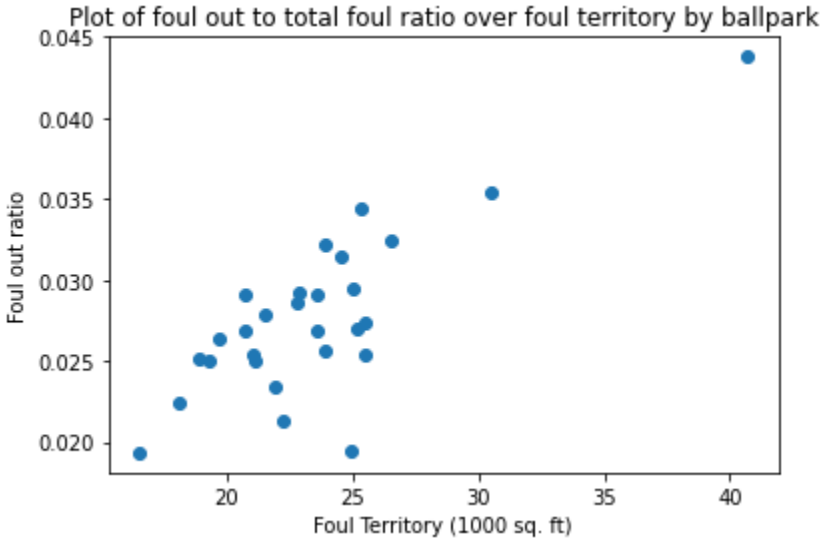


Figure 10 includes a scatter plot representing the ratio of foul outs to total foul balls by the amount of foul territory.

As seen in Figure 10, the scatter plots demonstrate a clear positive relationship between foul territory and the number of foul-outs that occur at that stadium, as well as the ratio of foul-outs to total balls hit in foul territory. For example, RingCentral Coliseum (home of the Oakland A's) has the highest amount of playable foul territory in the major leagues, with approximately 40,700 square feet of foul territory [2], and as shown by this relationship, has both the highest number of foul outs and the ratio of foul outs to total balls hit to foul territory. The increased amount of playable foul territory allows for the fielders to catch many balls in foul territory that in other stadiums would land in the stands, so that there is a distinct difference in the outcome of a play dependent completely on non-standardized field dimensions across different stadiums.

Home Run Distribution

Much like foul territory, the height and distance of the outfield walls in baseball are variable from the stadium to stadium, leading us to also analyze the distribution of home runs across different stadiums to examine any notable differences. To analyze the distribution of the locations of home runs, we first looked at all the home runs locations of the 2019 season. Through our data processing of the play-by-play data, each play that was a home run was tagged as 'HR' and where the location is (left-field, left-center, center field, right-center, and right-field) based on Retrosheet's hand-recorded hit location tags, which are limited to those five regions for accuracy purposes. After aggregating the total number of home runs by location for each stadium, we then analyzed the distribution of the home run locations per stadium.

Figure 10: Normalized and Non-Normalized Heatmaps of Home Run Locations per Stadium in the 2019 Season

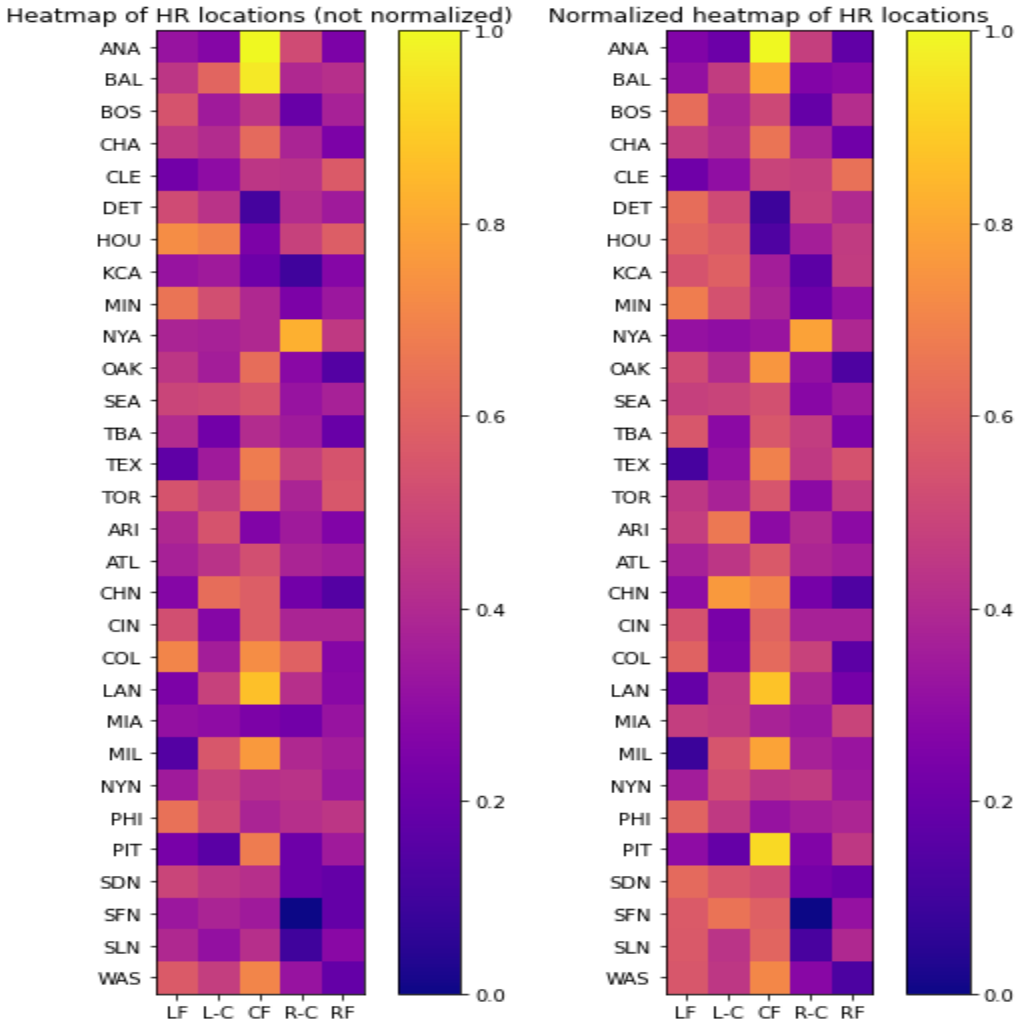


Figure 10 displays a produced heatmap of the aggregate home runs at each active stadium in the 2019 season by hit location, both unnormalized (left), where the color gradient is based on the aggregate number

of home runs hit to that location, and normalized (right), where the color gradient is based on the fraction of home runs hit in that stadium to that location. A more yellow square indicates for the right heatmap that more total home runs were hit to that location. For the left heatmap, a more yellow square indicates that a higher proportion of home runs hit at that stadium were hit at that location.

To note, we were not expecting to observe a uniform distribution at each stadium across each of the five locations as the general geometry of an outfield and the distribution of right-handed vs left-handed hitters would skew our results (this is expanded upon in Figure 11). As seen in Figure 10, the distribution of home runs to each of the five different hit locations varies drastically across stadiums. For example, Angels Stadium (home of the Los Angeles Angels) and Camden Yards (home of the Baltimore Orioles) both had some of the highest total numbers of home runs hit to center field, though when we examine the standardized distribution at each of those stadiums we note that the proportion of home runs hit to center field in Camden Yards is not as high compared to the rest of the league as the total volume. This lack of noticeably patterned distribution could be explained by the differing outfield dimensions; however, as we are dealing with a single-season heatmap, there are a couple of confounding factors to consider. Primarily, if the home team of a stadium's lineup in that season was mostly left-handed hitters, we would expect to see a higher number of home runs hit to right field, as baseball players hit home runs most often to their pull side. To account for this confounding factor, we expanded our dataset to include the 2010-2018 seasons and then ran the same analyses on our new dataset of 32 teams.

Figure 11: Aggregate Distribution of all Home Runs by Location in the 2010-2019 Seasons

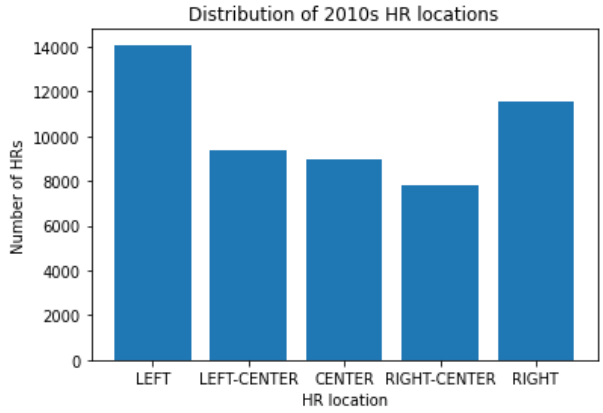


Figure 11 displays the Aggregate Home Run Location distribution for all 30 MLB teams across the 2010s by the total number of home runs to each of the five-hit locations defined by RetroSheet.

Figure 11 shows that the aggregate majority of home runs in the 2010s were hit to far left field and far-right field, with a higher proportion to left field. Hitters generate more power when they pull, meaning they change their swing to put slightly more time on the ball when swinging the bat, which directs the ball to the batter-side portion of the park.

Therefore, we would expect to see more home runs hit to left and right field, due to pulling the ball, and more generally to left field as there are more right-handed hitters than lefties in the majors, following general population distribution. We'd also expect fewer home runs hit to the center fields because the center-field wall is the furthest away from home plate, meaning the most power needs to be generated to hit a home run to this area.

Figure 12: Normalized and Non-Normalized Heatmaps of Home Run Locations per Stadium Aggregated Across the 2010-2019 Seasons

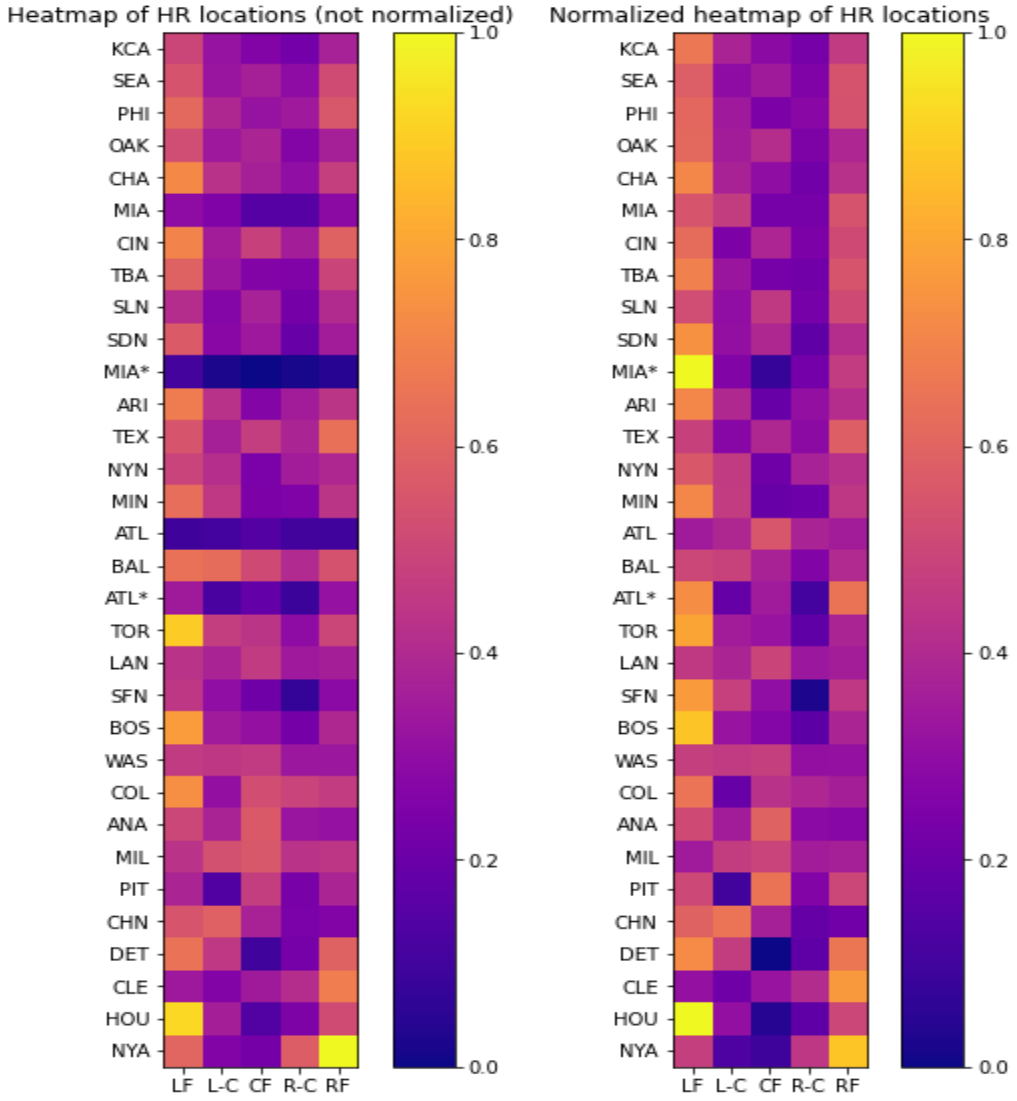


Figure 12 displays a produced heatmap of the aggregate home runs at each active stadium in the 2010-2019 season by hit location, both unnormalized (left), where the color gradient is based on the aggregate number of home runs hit to that location, and normalized (right), where the color gradient is based on the fraction of home runs hit in that stadium to that location. Both Atlanta and Miami switched stadiums in 2013 and 2017 respectively- here ATL* and MIA* refer to all games played in the old stadiums.

As seen in Figure 12, we're not able to effectively compare the non-standardized distributions due to the four stadiums associated with Atlanta and Miami not pulling from a full 10 seasons' worth of data, leading to the much lower frequency observed in the non-standardized heatmap. When we expand our heatmap to include all seasons and all 32 teams from the 2010-2019 seasons, we see that some stadiums have home-run standardized distributions that appear to match the general distribution found above. Some stadiums, however, like Cleveland and New York, don't seem to follow the league distribution when standardized. In order to determine how many teams follow the expected distribution of home run locations based on the aggregate distribution of home run locations in the league, we calculated a test statistic for a goodness-of-fit test, using the expected distribution derived from the aggregate distribution of the entire league, to account for the distribution in handedness across all major league players, as well as account for the confounding factor of some teams being built around more power-heavy hitters or more scrappy hitters.

Table 1: Chi-Squared Statistics for Home Run Location Distribution for select MLB Teams in the 2010-2019 Seasons

Stadium	Calculated Test Statistic Value	Accept or Reject Distribution at 95% Confidence ($t > 9.488$)	Accept or Reject Distribution at 99.5% Confidence ($t > 14.860$)
Kauffman Stadium (Kansas City Royals)	4.5931	**Accept**	**Accept**
T-Mobile Park (Seattle Mariners)	9.5841	Reject	**Accept**
Citizens Bank Park (Philadelphia Phillies)	13.0712	Reject	**Accept**
RingCentral Coliseum (Oakland A's)	13.5416	Reject	**Accept**
Guaranteed Rate Field (Chicago White Sox)	14.4533	Reject	**Accept**
…	…	…	**…**
Wrigley Field (Chicago Cubs)	169.3415	Reject	Reject
Comerica Park (Detroit Tigers)	194.4431	Reject	Reject
Progressive Field (Cleveland Guardians)	197.6663	Reject	Reject
Minute Maid Park	240.0251	Reject	Reject

(Houston Astros)			
Yankees Stadium (New York Yankees)	432.8489	Reject	Reject

Table 1 displays the lowest 5 and highest 5 test statistics for MLB Team Home Run Distributions over the 2010s, and their accept/reject status for the goodness-of-fit test based on said calculated test statistic at both 95% and 99.5% confidence. The expected value in our calculations was drawn from the overall MLB Home Run Location distribution, which can be found in Figure 13. The rest of this table can be found in Appendix A as Table 1.

Above, we can see that with 95% confidence, only the teams playing at Kauffman Stadium followed the expected distribution. With 99.5% confidence, the teams playing at Kauffman Stadium, T-Mobile Park, Citizens Bank Park, and RingCentral Coliseum followed the expected distribution; however, that is still only 12.5% of all stadiums we analyzed. On the other hand, the teams playing at Minute Maid Park and Yankee Stadium strayed the furthest from the expected distribution, with test statistic values so high there is no corresponding confidence level that would allow us to accept the null hypothesis that these two stadiums follow the major league distribution. We examined these two drastic outliers further below to illustrate

Figure 13: Aggregate Distributions of New York Yankees and Houston Astros Home Runs by Location in the 2010-2019 Seasons

Figure 13 displays the Home Run Location distributions for Minute Maid Stadium (right) and Yankee Stadium (left), aggregated over the 2010-2019 seasons. The calculated expected home run distribution based on the league aggregate distribution is overlaid to demonstrate each stadium's strong deviation away from the statistical expected distribution. The rest of the corresponding bar chart distributions for each of the 32 stadiums can be found in Appendix A as Figure 2.

To calculate the expected distribution for a team, we used the aggregate distribution displayed in Figure 12 to find the general probability of hitting a home run to each location, given that the hit is a home run. We then multiplied each probability by the total number of home runs hit at each stadium and used that as our expected distribution. Figure 12 demonstrates how our strongest deviators, Yankee Stadium and Minute Maid

Park, actually are different from the expected distribution. The Astros hit about 200 more home runs to left-field than expected across the 2010s, and the Yankees hit about 300 more home runs to right field than expected. While three of the five Home Run Locations appear to be very close to the expected at Minute Maid Park, all of the locations at Yankee Stadium appear to be very far off, leading to the extremely high test statistic value.

Specific Cases

2017 Houston Astros

In 2020, the MLB punished the Houston Astros for using an elaborate system to steal signs from opposing teams while playing at home during the 2017 regular season and playoffs. Their ability to steal signs allowed them to know what pitches were coming, which we hypothesized would significantly alter the team's hit and strikeout rates. Below, the expected rate for the Astros in a given year was calculated by finding the expected rate of teams that placed within 2 spots of the Astros in the overall standings. Specifically, hit rate was calculated by dividing the total number of bases acquired over the total number of at bats. The strikeout rate was calculated by dividing the number of strikeouts by the total number of at bats. Expected hit rate was calculated by taking the average hit rate of the teams that placed within 2 spots of the Astros in the year before.

$$Hit\ Rate\ =\ \frac{Singles\ +\ (Doubles * 2) + (Triples * 3) + (Home\ Runs * 4)}{Singles\ +\ (Doubles * 2) + (Triples * 3) + (Home\ Runs * 4) + Strikeouts + Outs + Walks}$$

$$Strikeout\ Rate\ =\ \frac{Strikeouts}{Singles\ +\ (Doubles * 2) + (Triples * 3) + (Home\ Runs * 4) + Strikeouts + Outs + Walks}$$

Figure 14: Houston Astros Hit and Strikeout Rates in the 2010-2019 Seasons

Figure 14 displays the Houston Astros' expected and real hit and strikeout rates by location and by year. The expected rates are the navy points, while the orange points represent the real rates. The team's rates at home are on the left while the right displays the away rates.

Figure 14 demonstrates a high-level view of the batting behavior of the Houston Astros in the 2010s. To be clear, a higher hit rate indicates better decision making at the plate, while a higher strikeout rate indicates worse decision making at the plate. This relates to sign stealing in that if the Astros batters know the pitches that are coming, they will likely make better decisions at the plate.

At home, the Astros consistently hit around their expected rate, except for 2015, 2017 and 2018. In 2015 and 2017 the Astros significantly outperformed expectations, while in 2018 they underperformed. During away games, the Astros consistently performed almost exactly as expected, except for the latter years when their overall performance as an organization rose to one of the best in the MLB. The lack of similar behavior at home and away demonstrates a level of home field advantage for the Astros, since they got better at home over time while also getting worse outside of Minute Maid Park.

In 2017 specifically, the Astros significantly outperformed their expected hit rate and strikeout rate. Diving deeper into the calculations for our expected rates, this indicates that either the teams that performed similarly to the Astros in previous years were all significantly worse than expected, or the Astros themselves had a massive jump in ability between 2016 and 2017. In 2016, the teams that placed around them were the Seattle Mariners, New York Yankees, Kansas City Royals and Miami Marlins. In 2017, all of

these teams placed similarly to what they had in 2016 except the Yankees who ended up losing to the Astros in the ALCS.

Because the Yankees and Astros had similar overall records between 2016 and 2019, and therefore similar placement in the standings, their expected rates follow a similar path to each other. Since the Astros and Yankees went on to face each other in the ALCS, their real rates should also be similar due to the fact that they were 2 of the 4 best teams in the league. The real and expected rates for the New York Yankees across the 2010s is displayed below in Figure 15.

Figure 15: New York Yankees Hit and Strikeout Rates in the 2010-2019 Seasons

Figure 15 displays the New York Yankees' expected and real hit and strikeout rates by location and by year. The expected rates are the navy points, while the gray points represent the real rates. The team's rates at home are on the left while the right displays the away rates.

Figure 15 demonstrates that in 2017, the Yankees practically performed the same as expected at home based on teams with similar performance in years before. The Yankees also performed similarly at home and away. This could be due to the fact that The Yankees' overall record across this timeframe was 921-699, meaning that throughout the entirety of this time they performed well above average. On the other hand, The Astros' record across the 2010s was 789-831. While their tanking in the early 2010s has an effect on the expected rates in later years, the purpose of including 'adjacent' teams in the expected rates was to account for this.

The Astros' rise in hit rate between 2016 and 2017 and their increased strikeout rate between 2017 and 2018 demonstrates that their 2017 performance at the plate was an anomaly. While we cannot statistically prove that the Astros cheated in 2017, the analysis

above demonstrates that their performance that year significantly differed from expectations and from their performance in adjacent years.

2018 Boston Red Sox

Later in 2020, the MLB punished the Boston Red Sox for stealing signs during their 2018 World Series run. While their acts of cheating differ from the Houston Astros', the impact on their performance should be somewhat similar. The main difference, however, is that their ability to cheat was not dependent on them being at home. Instead, the team used the MLB video replay system in order to relay signs to players, which can be done at any stadium by any team. While using the replay monitors to make in-game adjustments is allowed by the league, relaying signs to players based on video replay is not. A member of the Red Sox organization sat in the video replay centers of each stadium and relayed signs to players in real time: a clear violation of MLB policies. In order to demonstrate how egregious their change in performance was, we performed a similar analysis to what we did on the 2017 Houston Astros.

Since the cheating was done both at home and away, we should not expect the cheating to be apparent in the Red Sox's home-away splits. However, it is entirely possible that their ability to cheat using the video replay monitors was better at home since they were more comfortable with the stadium's facilities.

Using the expected rates by year is crucial in order to show how the Red Sox significantly outperformed similar teams in 2018 based on past performance. While they won the World Series that year, and many of their players made noticeable jumps in performance, we cannot prove that the Red Sox's better performance is independent of their cheating in 2018.

Therefore, we expected the 2018 Red Sox to significantly outperform their previous seasons, as well as their expected performance for that year. In 2017, the teams that placed similarly to the Red Sox were the Houston Astros, Washington Nationals, Arizona Diamondbacks and Chicago Cubs. All of these teams had strong seasons in 2018 except for the Arizona Diamondbacks, indicating that the expected rate for the 2018 Boston Red Sox should not deviate too much from their real performance.

Figure 16: Boston Red Sox Hit and Strikeout Rates in the 2010-2019 Seasons

Figure 16 displays the Boston Red Sox's expected and real hit and strikeout rates by location and by year. The expected rates are the black points, while the red points represent the real rates. The team's rates at home are on the left while the right displays the away rates.

Figure 16 displays the Boston Red Sox expected and real performance at the plate throughout the 2010s. Interestingly, their real hit rate at home deviates consistently from their expected hit rate, except for in 2017 where they underperformed significantly. In 2018, their real hit rate jumped back up to about the same level as their rates in years before. Similarly, the Red Sox's strikeout rate at home somewhat matched expectations across the 2010s.

Lessons Learned

The main lesson we learned from this project was the amount of time needed to be dedicated to processing the data that was gathered. While at first glance, plays in baseball may seem intuitive with only a few types of plays, the number of combinations possible for each play was more than anticipated, and drastically increased the complexity. Especially when processing the play-by-play data, we initially wanted to make a generic function to categorize each play but ran into numerous use cases that differed only slightly from previous ones found. This became apparent when running the function for different teams and different seasons where another case had to be added.

Another lesson is that sometimes a dataset can't be represented in low-dimensional space. We attempted to run PCA many times over, dropping different columns in an attempt to derive a low-dimensional space to work in, but could never get a satisfactory percentage of the variance explained by a small enough number of features.

Finally, as we created this report we realized that we needed to be thoughtful with the visualizations we created, since we will likely be addressing readers with a range of statistics and/or baseball knowledge. In reports such as this, it is important to take the time to create a clear narrative of how we got from collecting the data to our results.

Conclusions

High-Level Home Field Advantage

Upon examining our results for the offensive performance differential between the home team and the away team over the last ten seasons of baseball, we can observe and conclude several things. Firstly, as seen in Figure 4, there is a clear difference in how teams perform offensively in terms of total runs scored when they play at their home stadium vs on the road. However, that difference swings in both directions; some teams in the last ten seasons have performed better on the road, and some teams have performed better at home, so we cannot conclude that there is a general advantage to be had from simply playing at the home stadium. When we examine the differences in offensive performance between the home team and the away team at the same stadium, we observe on average that the run differential per game between the home team and the away team is approximately 0, indicating that neither the home nor the away team scores more runs than the other at all stadiums. We cannot conclude that there is any inherent offensive performance boost for the home team that would constitute an advantage over the away team at the same stadium

Stadium Differential Advantage

Upon examining our results for offensive performance differential between teams playing at different stadiums, several conclusions come to light. Firstly, the non-standardization of baseball dimensions, specifically the differences in the amount of playable foul territory and outfield wall height and differences. In the case of differences in playable foul territory, the positive correlation between foul territory and foul outs could be constituted as an advantage for the home team. An increase in foul outs at a certain stadium means that the pitchers at that stadium are benefitting from foul outs that in other stadiums land in the stands, and therefore the pitchers of the home team benefit from these additional outs for half the games they pitch over the course of a whole season, which could be considered an advantage for the home team over other teams playing at their

home stadiums. When it comes to examining the differences in outfield dimensions, we note a few key points. We would expect to see some variation in distribution of home run locations based on the different fence heights and distances from home plate. The extent of differences in distribution is especially notable because there is almost no general distribution of location we can extrapolate from- almost every stadium has a unique distribution. We can't necessarily conclude that the differences in distribution stem from the differences in outfield wall height and distance from home plate. However, we can conclude that each stadium possessing its own distribution could constitute an advantage for the home team if the home team specifically targets hitters whose own hit patterns match that of the stadium distribution; i.e. Yankees Stadium favors right field due to the right field wall being closer to home plate than other stadiums, so if the Yankees were to build a team of left-handed hitters who favor right field, that would be an advantage to them playing at their home stadium.

Specific Cases

Using similar methods to analyze stadium-based advantages, we were able to demonstrate how advantageous the Astros' cheating in 2017 really was. Through this process, we were able to discern general behaviors of MLB teams across long periods of time. For example, using the adjacent team method of calculating expected hit and strikeout rates was highly accurate across most MLB teams. In calculating expected rates, the inclusion of teams that went from 'worst-to-first' such as the 2013 Arizona Diamondbacks does not have a significant influence because of the volume of data on each team used in each calculation. The use of play-by-play data, which contains specific information on every play of every game, only allows for major changes in expectations if an entire team's behavior at the plate changes. Baseball games have at least nine different batters every game, all of whom have matured into certain behaviors in game that are unlikely to change, especially as a team, season by season. This allows us, as researchers,

Appendix

Appendix A: Additional Tables and Figures

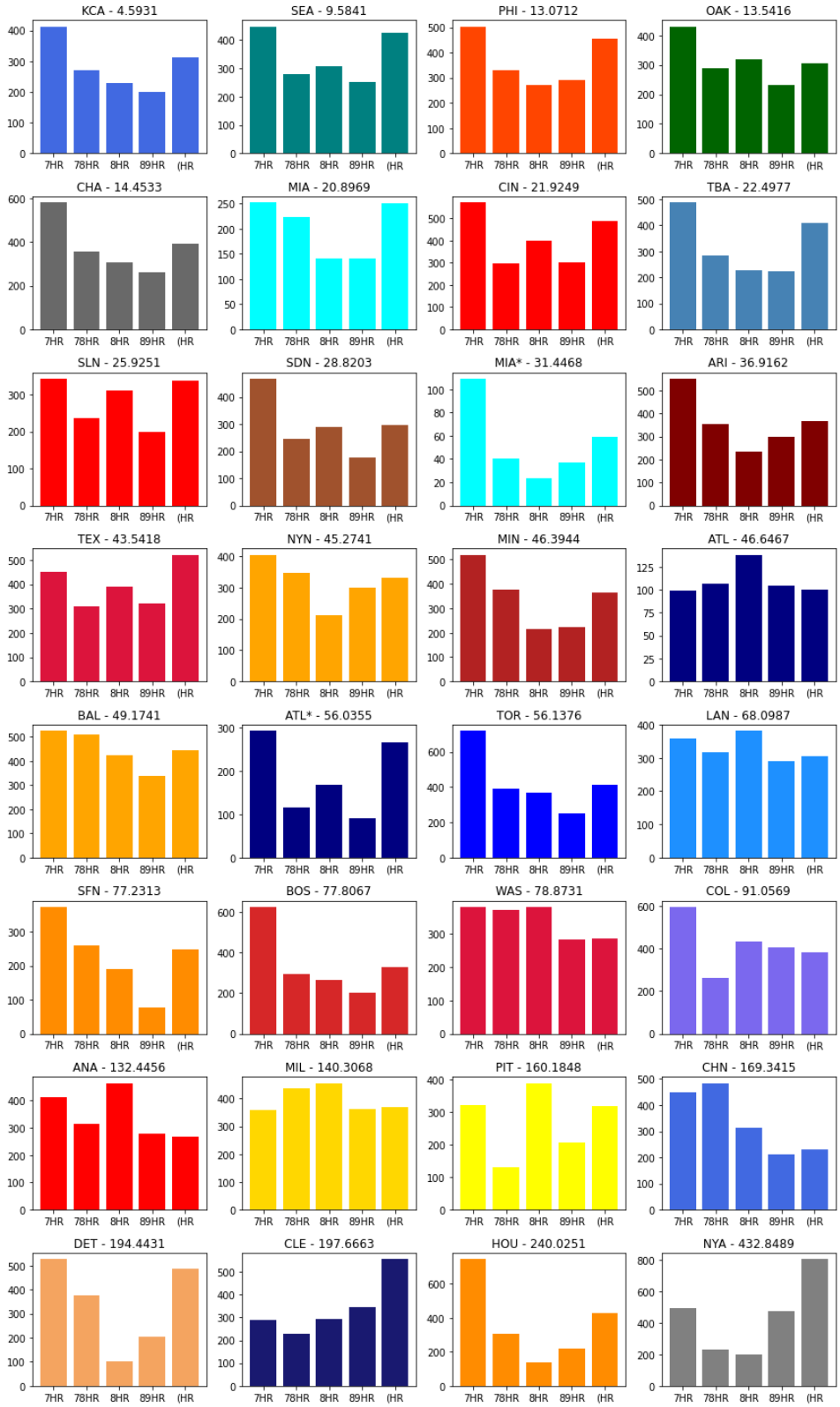
Table 1: Calculated Chi-Squared Test Statistics for Homerun Distribution for each MLB Stadium

Stadium	Calculated Test Statistic Value	Accept or Reject Distribution at 95% Confidence ($t > 9.488$)	Accept or Reject Distribution at 99.5% Confidence ($t > 14.860$)
Kauffman Stadium (Kansas City Royals)	4.5931	**Accept**	**Accept**
T-Mobile Park (Seattle Mariners)	9.5841	Reject	**Accept**
Citizens Bank Park (Philadelphia Phillies)	13.0712	Reject	**Accept**
RingCentral Coliseum (Oakland A's)	13.5416	Reject	**Accept**
Guaranteed Rate Field (Chicago White Sox)	14.4533	Reject	**Accept**
loanDepot Park (Miami Marlins (2012-2019))	20.8969	Reject	Reject
Great American Ball Park (Cincinnati Reds)	21.9249	Reject	Reject
Tropicana Field (Tampa Bay Rays)	22.4977	Reject	Reject
Busch Stadium (St. Louis Cardinals)	25.9251	Reject	Reject
Petco Park (San Diego Padres)	28.8203	Reject	Reject
Sun Life Stadium (Florida Marlins (2010-2011))	31.4468	Reject	Reject
Chase Field (Arizona Diamondbacks)	36.9162	Reject	Reject
Rangers Ballpark	43.5418	Reject	Reject

(Texas Rangers)			
Citi Field (New York Mets)	45.2741	Reject	Reject
Target Field (Minnesota Twins)	46.3944	Reject	Reject
Truist Park (Atlanta Braves (2017-2019))	46.6467	Reject	Reject
Oriole Park at Camden Yards (Baltimore Orioles)	49.1741	Reject	Reject
Turner Field (Atlanta Braves (2010-2016))	56.0355	Reject	Reject
Rogers Center (Toronto Blue Jays)	56.1376	Reject	Reject
Dodgers Stadium (Los Angeles Dodgers)	68.0987	Reject	Reject
Oracle Park (San Francisco Giants)	77.2313	Reject	Reject
Fenway Park (Boston Red Sox)	77.8067	Reject	Reject
Nationals Park (Washington Nationals)	78.8731	Reject	Reject
Coors Field (Colorado Rockies)	91.0569	Reject	Reject
Angel Stadium (Los Angeles Angels)	132.4456	Reject	Reject
American Family Field (Milwaukee Brewers)	140.3068	Reject	Reject
PNC Park (Pittsburgh Pirates)	160.1848	Reject	Reject
Wrigley Field (Chicago Cubs)	169.3415	Reject	Reject
Comerica Park (Detroit Tigers)	194.4431	Reject	Reject
Progressive Field (Cleveland Guardians)	197.6663	Reject	Reject

Minute Maid Park (Houston Astros)	240.0251	Reject	Reject
Yankees Stadium (New York Yankees)	432.8489	Reject	Reject

Figure 2: Home Run Location Distributions for all 30 MLB Teams across the 2010-2019 Seasons



References:

[1] The information used here was obtained free of charge and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at "www.retrosheet.org".
Retrosheet Game Logs, https://www.retrosheet.org/gamelogs/index.html.

[2] The information used here was obtained free of charge and is copyrighted by Clem's Baseball Blog. Interested parties may contact Andrew Clem at "http://www.andrewclem.com/Baseball.php".
Stadium Lists by name and by city, http://www.andrewclem.com/Baseball/Stadium_lists.html