Style of Play and Attendance in MLB

Margot Douillet

March 18, 2021

Introduction

Before COVID-19 changed the way in which live sports were played, Major League Baseball was already facing a crisis. In 2018, MLB saw its average attendance drop to the lowest mark in 15 years [1], and this has been accompanied by a steady increase in the average age of MLB fans, up to 57 years old in 2019 [2]. There have been many sources of blame for giving baseball a reputation of monotony, among them changes in the style of play and an increase in the average duration of MLB games.

Within baseball, "playing style" is a colloquial term used to describe either a team's pitching tendencies, hitting tendencies, or a combination of both. Teams with high stolen base and home runs totals are considered to have a "flashier", more exciting style of play, and thus may be more capable of drawing fans. The connection between length of games and styles of play is somewhat tenuous. Because strikeouts require more pitches than ground ball or fly ball outs, and walks require more pitches than hits, teams with higher strikeout and walk totals (both in terms of hitting) would theoretically play longer games [9]. There does appear to be a correlation between MLB's increase in strikeouts and decrease in attendance, although this correlation does not imply a causational relationship. It may be, then, that different styles of play have a hand in determining a team's attendance - particularly styles of play which would produce longer games.

This study aims to examine whether different styles of play were associated with higher attendance during the 2019 season - specifically if faster-paced play might have boosted ticket sales. There is a substantial body of research linking more team wins to higher attendance, such as a 2009 study by Michael C. Davis, which found that each additional win above a 0.500 winning percentage adds between 138 and 380 fans per game [3]. Other factors have also been found to contribute to ticket sales. In 1999, Michael R. Butler found that matchups between teams in different leagues led to about a 7% increase in attendance [4]. Denaux et. al., 2011, found that the date, fan interest, and the team's success all factored into attendance [5]. Mark MacDonald (2000), even found that attendance could be partially explained by ballpark promotions [6].

There is no significant research done, however, on the affect of a team's playing style on its attendance.

In order to categorize playing style, we will use k-means clustering to separate Major League Baseball teams into two different, distinct clusters based on various statistics for team performance. First, we will use team hitting statistics to create two clusters, then we will use pitching statistics to form the clusters. Finally, we will use a combination of pitching and hitting statistics to separate MLB teams into two groups based on whether their pitching and hitting tendencies are theoretically related to longer games. For each of these three splits, we will compare the average and road attendance of teams between clusters using bootstrapped confidence intervals, bootstrapped p-values, and distribution-free permutation p-values.

Data

Attendance data comes from ESPN.com, and all hitting and pitching metrics come from MLB.com. The original data was aggregated and cleaned to remove unecessary variables. The primary attendance statistics being used are teams' average 2019 attendance, and teams' average attendance in 2019 for road games. Fans may choose what games to see based on the visitng team more than the home team, which makes road attendance data more interesting than home attendance data.

Due to the evidence suggesting a correlation between wins and attendance, average and average road attendance figures were adjusted based on a team's win percentage, and then robustly normalized. In order to perform this adjustment, team wins were first normalized. Then the average and road average attendance figures were divided by the percentiles of team wins. Afterwards, these adjusted average and road average attendance figures were robustly normalized. Robust normalization is calculated by subtracting the median, and dividing by the inter-quartile range [7]. It tends to be more robust against outliers. Hereafter, the robustly normalized adjusted average and road average attendance figures will merely be referred to as "average attendance" and "average road attendance".

Following robust normalization, there appeared to be several outliers. Rosner's test, which can test multiple outliers at once, was run, and consequentially three observations were removed. Unlike other outlier tests, Rosner's test tends to avoid the problem of masking, in which two close outliers go undetected [8]. A fourth observation was determined to be an outlier in average road attendance figures, but not in average attendance figures, and so was not removed from the dataset.

There is little correlation between the three hitting metrics of interest: walks, stolen bases, and strikeouts. Table 1 shows a correlation matrix for the three hitting statistics of interest. Strikeouts and walks have a negative correlation, suggesting that teams with high walk totals tend to strike out less.

Table 2 shows the correlation matrix for pitching statistics. For pitching metrics, there is a very high

	Walks	Stolen Bases	Strikeouts
Walks	-	0.213	-0.281
Stolen Bases	0.213	-	0.190
Strikeouts	-0.281	0.190	-

Table 1: Correlation matrix for the three hitting statistics of interest

correlation between the three predictors of interest. Walks and Walks and Hits per Inning Pitched (WHIP) have a correlation of 0.989, which is to be expected because walks are included in the computation of WHIP.

	Walks	Strikeouts	WHIP
Walks	-	0.921	0.989
Strikeouts	0.921	-	0.939
WHIP	0.989	0.939	-

Table 2: Correlation matrix for the three hitting statistics of interest

The relationship between the three pitching metrics of interest is shown in figure 1 below. The x-axis has walks allowed by pitchers, the y-axis is strikeouts by pitchers, and the color of the points corresponds to a pitcher's Walks and Hits per Inning Pitched (WHIP). There are two very distinct groups in the plot - one with low walks, low strikeouts, and low WHIP.

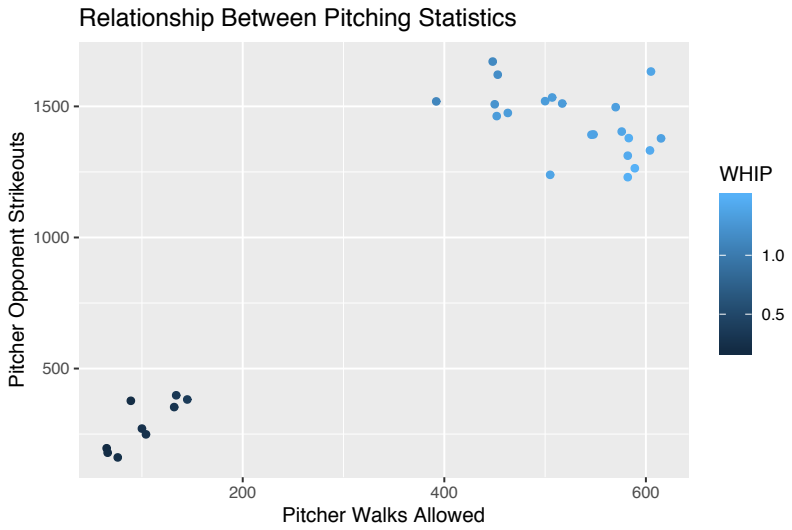

Figure 1: Relationship between mostly highly correlated pitching statistics

The distribution of average attendance (after the adjustments previously mentioned) is shown in figure 2 below. The distribution does appear slightly right-skewed, with a few teams having high average attendance figures.

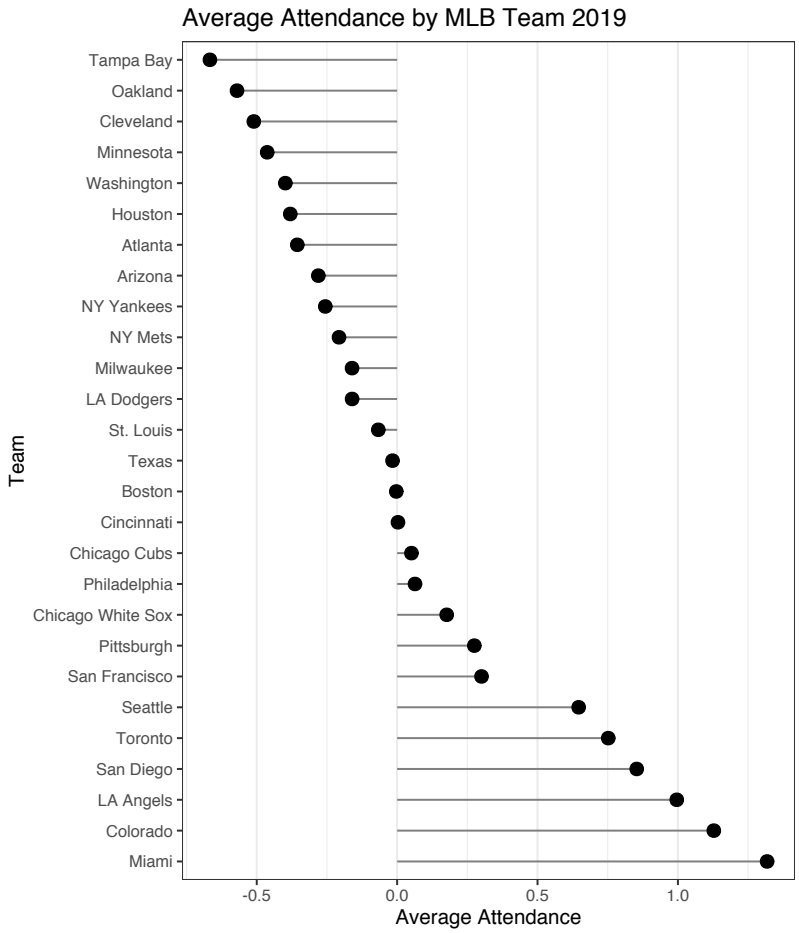

Figure 2: Normalized average attendance, adjusted by wins, for each MLB team during the 2019 season

Methods

K-means clustering is an unsupervised learning algorithm. It partitions objects into K distinct, non-overlapping clusters. Using j predictors, the K-means algorithm first randomly assigns observations to one of K clusters. It then computes the cluster centroid - in this case the vector of predictor means - and re-assigns each observation to the cluster with the nearest centroid. These last two steps are repeated until the within-cluster variation - here defined as the sum Euclidean distances between predictors - is minimized. Here we will direct the algorithm to produce two clusters, in part due to the small sample size.

The k-means clustering will first be performed using measures of teams' pitching metrics. The three predictors used to create the clusters will be teams' strikeouts, saves, and Walks and Hits per Inning Pitched (WHIP). A save is credited to a pitcher who enters a game in the late innings and preserves their team's lead (for more on the circumstances required to earn a "save", see mlb.com's glossary of statistics). Saves are typically used to quantify the strength of a team's bullpen, or their corp of relief pitchers - pitchers who

do not start the game. WHIP represents the number of walks and hits a pitcher allows per inning of work.

The second clustering algorithm will be run using teams' hitting metrics. Walks, stolen bases, and strikeouts will be used as the three predictors to form the two clusters. Stolen bases have fallen out of favor in recent years, in part due to the "Moneyball" style of play introduced in the early 2000s, which features low stolen bases and strikeout totals, and high walk rates.

Finally, the last clustering algorithm will be run on both pitching and hitting metrics. It will split MLB teams into two groups based on teams' hitters' hits and walks, as well as their pitchers' strikeouts and hits allowed. Because walks and strikeouts require more pitches than ground ball and fly ball outs, these would theoretically be associated with longer games. Theoretically, higher hit totals would lead to more runners on base, and therefore longer innings too.

For each of the three clustering algorithms, we will compare the mean difference in average attendance and the mean difference in average road attendance between the two clusters. This comparison will be done using both bootstrapped percentile confidence intervals and bootstrapped p-values. Bootstrapping refers to repeatedly drawing from the sample distributions with replacement. We must assume that the sample is sufficiently large to draw from replacement, and it represents a simple random sample. The advantages of bootstrapping are that it does not require assumptions of normality, and does not require the same complicated calculations as parametric methods.

All bootstraps will draw $B = 1000$ observations from the original data, and compute the confidence intervals with these samples. For each iteration of the bootstrap, we will compute $\bar{X}^*_{AO} - \bar{X}^*_{BO}$, which is the difference of average attendance between group A and group B (determined by clustering), and $\bar{X}^*_{AR} - \bar{X}^*_{BR}$, which is the difference of average road attendance between group A and group B. The 95% confidence interval for each of these differences is the 2.5% and 97.5% quantiles of the bootstrapped differences.

For $H_0 : \mu_A < \mu_B$, that is group A has a lower mean attendance than group B, bootstrapped p-values are calculated by taking the proportion of bootsrapped mean differences that are less than 0. For $H_0 : \mu_A > \mu_B$, bootstrapped p-values are calculated by taking the proportion of bootsrapped mean differences that are greater than 0.

While we assume that the average attendance and road attendance figures for both teams are approximately normally distributed, this is not necessarily the case. As a result, we will also perform a permutation test on the difference in average attendance and difference in average road attendance for each of the three clustering iterations. In order to perform a permutation exact test, we will first calculate the difference in mean ranks of average attendance between group A and group B. This is treated as the null test statistic. Then we will permute only the group variable (the labels designated by clustering) 1,000 times, each time calculating the difference in mean ranks of average attendance between group A and group B. The p-value

will be the proportion of differences that "more extreme" than the null test statistic. For $H_0 : \mu_A > \mu_B$ the p-value will be the proportion of differences that are greater than the null hypothesis.

A permutation test is based on the ranks of the observations, and therefore does not require the same assumptions about the underlying distribution of the data that a t-test requires (namely that the data is normally distributed). It also does not require the sample size to be very large. The permutation test does require the assumption that all of our permuted samples are equally likely to occur.

Simulation

To assess the accuracy of clustering, we created a simulated dataset using the three pitching metrics of interest from the MLB dataset. Our simulated data contained a WHIP variable, Strikeouts variable, and Saves variable. The first and second group's predictors were selected from normal distributions with the same standard deviations. We bootstrapped the simulated data, performed clustering, and calculated the classification error for the resulting clustering label. This process was repeated 1,000 times for each simulated dataset.

We first selected the first and second group's predictors so their means were three standard deviations apart. This was meant to simulate two groups that were very distinct. Then this process was repeated, this time using simulated data in which the means for the predictors were one standard deviation apart. This was meant to represent clusters which were only slightly different. Finally, we simulated a dataset in which observations either belonged to the first or second group, but their predictors were drawn from the same distribution. This was meant to represent a dataset in which there is virtually no separation between the groups, and cluster labels are essentially random.

Figure 3 shows the distribution of the cluster classification error for each of the three simulated datasets. Note that for the different clusters and slightly different clusters, the distributions of errors appear to me bimodal. This is likely because the algorithm identified a delinieation between the two clusters, but assigned the wrong cluster labels. Taking this into account, it appears that clustering is far more accurate on datasets in which there is a clear delination between two groups.

The bootstrapped p-values and confidence intervals computed in this paper will all be based on bootstrapping - repeated sampling with replacement. Bootstrapping relies on the central limit theorem, which states that the bootstrapped sampling distribution should approximate the population distribution. However, this relies on the assumption that the original sample $X_1, ..., X_n$ from which we are resampling $X_1^*, ..., X_n^*$ is a simple random sample. That is, we assume that each $X_1, ..., X_n$ had an equal probability of being drawn from the population distribution.

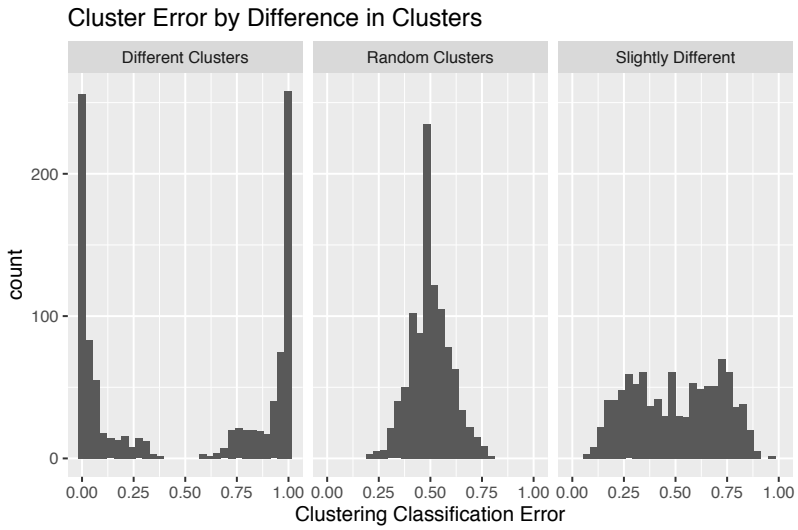

Figure 3: Distribution of clustering classification error for three separation scenarios

Consider a population distribution that is uniform on -10, 10. A simple random sample (unbiased) of size 100 can be drawn from this distribution with replacement, and the mean of this sample can be computed. We can also draw a biased sample of size 100, with replacement, from this distribution, where the probability of each point being sampled can be expressed by the exponential density (meaning positive values have a higher probability of being drawn). If we bootstrap - i.e. repeat this process 1,000 times, then this should approximate the population mean.

The distribution of the bootstrapped sample means from the biased sample and unbiased sample, respectively, are shown in figure 2. If the population distribution is uniform on -10, 10, then we know that the true mean is 0.05. The biased sample does a poor job of estimating the population mean. The distribution of the bootstrapped sample mean from the biased sample does not contain the true parameter. The distribution of the bootstrapped sample mean from the unbiased sample on the other hand appears to be centered around the true mean.

We can see that the manner in which a sample is drawn from the original population has an impact on its ability to estimate parameters. Simple random samples are far more accurate in predicting true parameters than biased samples. Even if the sample is randomly drawn, it must also be large enough. While bootstrapping is intended to be used for sample sizes that are not very large, it is not a magic solution. If the sample used for bootstrapping is small to begin with, it can produce an inaccurate estimate of parameters.

Consider a population that follows the standard normal distribution, with mean 0 and standard deviation 1, and a sample of 10 observations drawn from this population. The sample size is so small that it doesn't matter how many bootstrapped iterations are performed, the parameter estimates will still be biased. Fig-

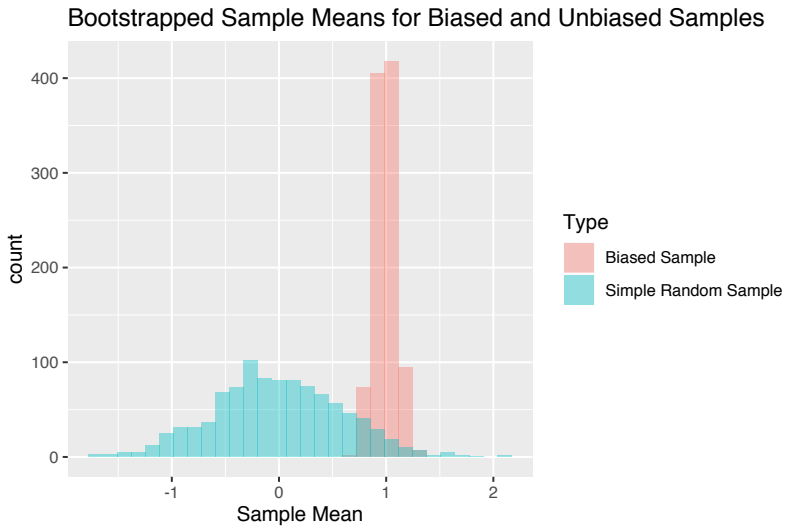

Figure 4: Distribution of the sample means from a biased sample and an unbiased sample after bootstrapping

ure 3 shows the distribution of bootstrapped sample means from this small sample after 1,000 bootstrap iterations. The vertical blue line represents the "true" mean of 0.

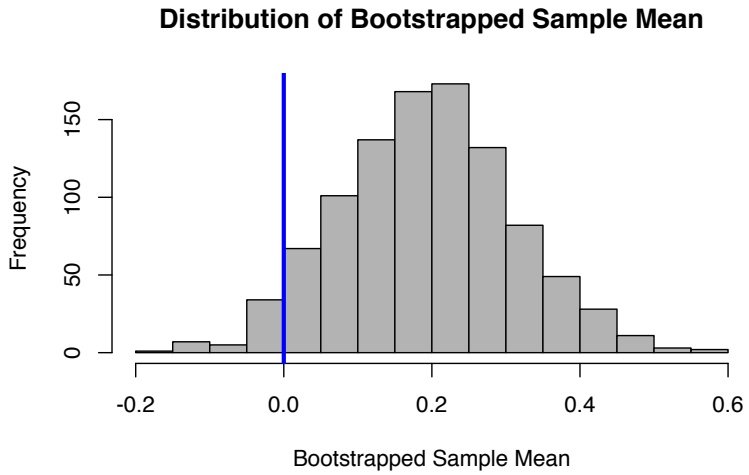

Figure 5: Distribution of bootstrapped sample mean from a sample size of n = 10. The blue line represents the "true" mean of 0

Results

The centers for the first two clusters, based on Pitching statistics, are shown below. The first group will be referred to as the "Contact" group, since these teams have low strikeout, walk, and hit totals. The second group will be referred to as the "Power" group, as its teams have high strikeout, walk, and hit totals, as well

as a strong bullpen.

Group	Strikeouts	WHIP	Saves
Contact	318.000	0.273	8.571
Power	1452.250	1.324	39.350

Table 3: Centroids for first two cluster, based on pitching statistics

A 3D scatterplot of the two different groups of teams, separated based on pitching metrics, under the first clustering algorithm, are shown in figure 2. The two different clusters appear particularly distinct.

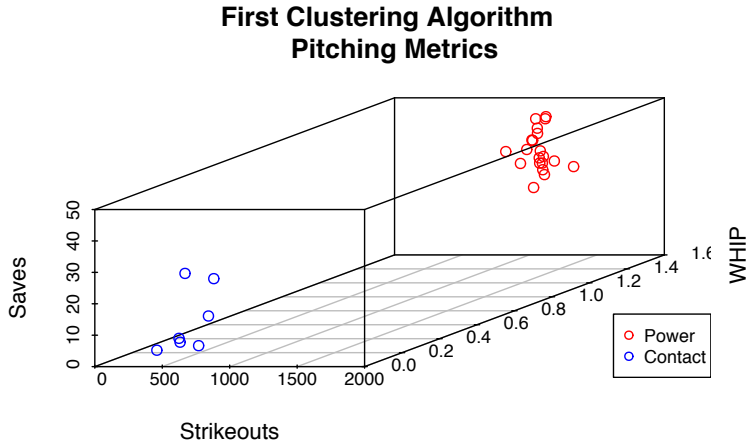

Figure 6: 3D scatterplot of the first clusters, based on pitching metrics

The difference in mean attendance between the average Contact team and the average Power team was -0.159. The mean difference in average road attendance between these two groups was -0.133. Both of these figures would suggest that Power teams tend to draw more fans. The 95% bootstrapped confidence interval for the mean difference in average attendance was (-0.523, 0.180). The 95% bootstrapped confidence interval for the mean difference in average road attendance was (-0.659, 0.370).

Both of these confidence intervals contain 0, and the bootstrapped p-values reflect this. The null hypothesis was $H_0 = \mu_C - \mu_P \leq 0$, where μ_P is the mean average attendance for the Power group, and μ_C is the mean average attendance for the Control group. For the average attendance, the bootstrapped p-value for a difference in means was 0.194. Testing this null hypothesis for mean average road attendance yielded a bootstrapped p-value of 0.323. For the permutation test on average attendance, the null test statistic was 0.193. The probability of observing a test statistic at least as extreme - the permutation p-value - was 0.509. For average road attendance, the permutation test statistic was -0.193, with corresponding p-value 0.292.

The cluster centroids for the second clustering algorithm, based on hitting metrics, is shown in table 2.

Group	Walks	Stolen Bases	Strikeouts
Moneyball	559.909	69.182	1313.545
Modern	519.167	81.222	1487.722

Table 4: Centroids for second two clusters, based on hitting metrics

A plot of the two separate clusters is shown in figure 3 below. This delineation appears less clear than the first, with the two groups appearing to have fairly close centroids, and higher within-cluster variation than the clusters separated by pitching metrics.

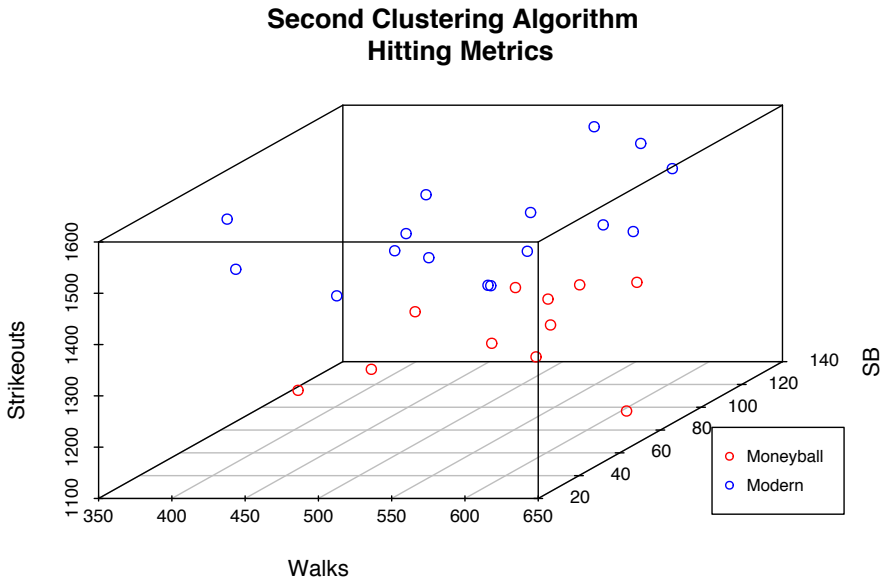

Figure 7: 3D scatterplot of the second clusters, based on hitting metrics

The first group will be referred to as the "Moneyball" group, characterized by an old-school mixture of high walk totals, low stolen bases, and low strikeouts. The second, "Modern", group represents lower walk totals, more stolen bases, and higher strikeout numbers. These two groups appear more distinct than those separated based on hitting metrics. The mean difference in average attendance was -0.390, while the average difference in average road attendance was -0.503. These would both suggest that teams with a more Modern style of play attract more fans.

The 95% percentile bootstrapped confidence interval for the mean difference in average attendance was (-0.523, 0.180,), and the 95% percentile bootstrapped confidence interval for the mean difference in average road attendance was (0.659, -0.370). Testing the null hypothesis $H_0 = \mu_{MB} - \mu_{MD} \geq 0$, where μ_{MB} is the mean average attendance for the "Moneyball" group, and μ_{MD} is the mean average attendance for the

Modern group. The bootstrapped p-value for the mean difference in average attendance was 0.022, and the bootstrapped p-value for the mean difference in average road attendance was 0.014. Both are statistically significant at the 5% level. For average attendance, the permutation test statistic was -6.75, with p-value 0.008. For road attendance, the permutation test statistic was -5.98 with p-value 0.026.

The final cluster separated teams based on a mixture of pitching and hitting metrics: Hits (hitting), walks (hitting), strikeouts (pitching), and hits (pitching). Table 3 below shows the centroids for the two different clusters. The first group is characterized by poor offense - low hit and walk totals for its hitters - and low strikeout and hit totals for its pitchers. Because walks, hits and strikeouts are theoretically associated with more pitches, and thus longer innings, the first group is referred to as the "Fast" group, and the second as the "Slow" group.

Group	Hits	Walks	Strikeouts (pitching)	Hits (pitching)
Fast	1381.625	514.250	298.375	284.125
Slow	1412.048	542.381	1441.667	1398.190

Table 5: Centroids for the last two clusters, based on hitting and pitching metrics

The mean difference in average attendance between the two groups was 0.159, with 95% bootstrapped confidence interval (-0.145, 0.503). The mean difference in average road attendance between the two groups was -0.133, with 95% bootstrapped confidence interval was (-0.352, 0.705). Both confidence intervals contain 0, suggesting there is no evidence that the true difference in attendance between fast teams and slow teams is greater than 0.

Testing the null hypothesis $H_0 = \mu_F - \mu_S \geq 0$, where μ_F is the mean average attendance for the Fast group, and μ_S is the mean average attendance for the Slow group yielded a bootstrapped p-value of 0.178. Testing this same null hypothesis for the mean difference in average road attendance yielded a bootstrapped p-value of 0.301. The permutation p-value for the difference in average attendance was 0.167, while the permutation p-value for the difference in average road attendance was 0.338. Neither of these are statistically significant at the 5% level. The null permutation test statistic for average attendance was -0.193, with corresponding p-value 0.483. The permutation null test statistic for average road attendance was 1.928, with p-value 0.293.

Figure 4 below shows the distribution of average attendance of each team, faceted by their group (Fast vs. Slow) under the last clustering algorithm.

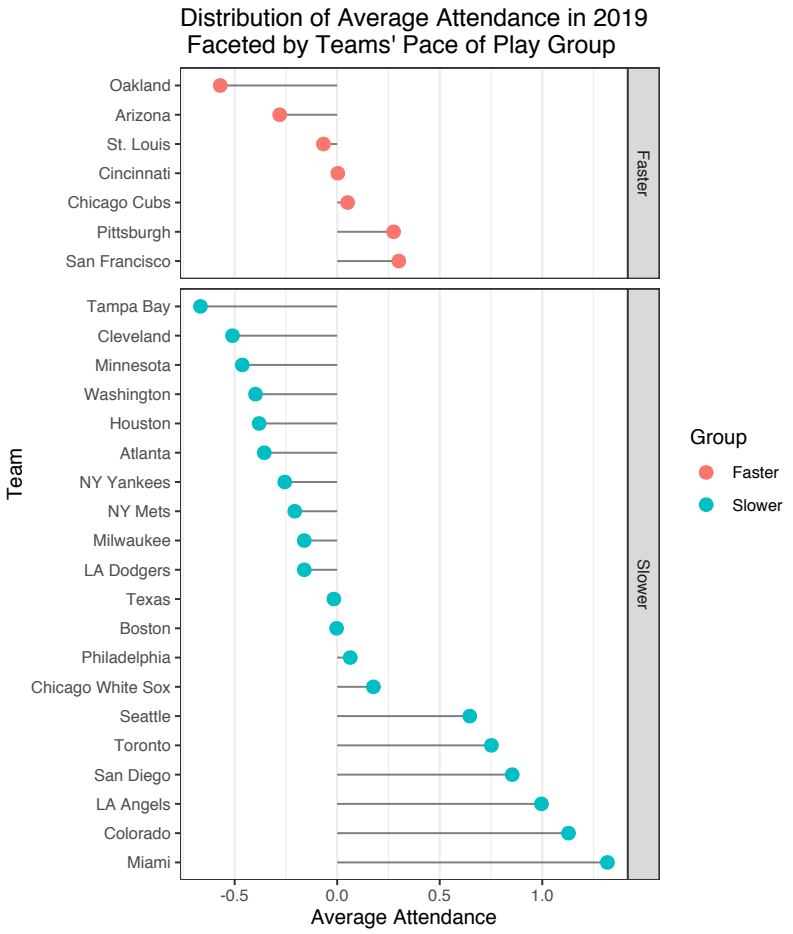

Figure 8: Distribution of average attendance between teams with a slow pace of play and teams with a fast pace of play

Discussion

The first clusters we made were based on pitching statistics. While the two clusters appeared very dissimilar, our two clusters, Contact and Power teams did not have a statistically significant negative mean difference in average attendance or negative mean difference in average road attendance at the 5% level ($p > 0.05$) based on either the bootstrapped p-values or permutation tests. It appears then, that a difference in pitching styles does not have an effect on attendance.

Next, we separated teams based on hitting statistics. The two groups, "Moneyball" teams and "Modern" teams did not appear very dissimilar, but the mean difference in average attendance and mean difference in average road attendance were both statistically significantly positive at the 5% level ($p < 0.05$), based on both the bootstrapped p-values and permutation tests. Therefore, there is sufficient evidence to conclude that teams with a more Modern hitting style have higher attendance overall, and on the road. This "Modern" hitting style was characterized in the clustering algorithm by lower walk totals, higher stolen bases, and more

strikeouts. While the Modern group had higher strikeout totals, they also had lower walks totals. It does not appear that there would be any significant difference in pace of play between these two types of teams.

Finally, we separated teams based on theoretical paces of play. The Fast group was characterized by hitters getting more hits and walks, pitchers getting more strikeouts, and allowing more hits. There was no statistically significant positive mean difference in average attendance or positive mean difference in average road attendance at the 5% level ($p > 0.05$), based on either the bootstrapped p-values and the permutation tests.

The pace of play demarcations, "Fast" and "Slow", are only based on theoretical associations between certain pitching and hitting metrics and time. It is possible that the associations between metrics and pace of play assumed in this study are wrong, and therefore the results have been biased. It is also possible that there is an actual correlation between the metrics considered and pace of play, but that a causational relationship does not exist.

While the k-means clustering appeared to separate fairly distinct clusters, the final clusters do depend somewhat on the initial random assignment to clusters. In most cases, clustering is run several times to select the best clustering delineation. This was not done for the purposes of this study since the sample size was so small to begin with, there would be little (if any) difference in the clusters if run another time. In the simulation section, we saw that clustering tends to perform better on data that can be clearly separated into k groups. While our pitching data could be very clearly separated into two clusters, the hitting data was separated into less distinct clusters.

It is difficult to repeat this study with additional data, since attendance figures can vary wildly from one year to the next. The small sample size was somewhat circumnavigated through bootstrapping, but still could bring the results into question. We saw that a small sample can bias the results of bootstrapping. In addition, we assumed that the data was a simple random sample - which we saw had a major impact on the accuracy of bootstrapping - which may not be true.

We initially identified that wins were heavily correlated with attendance, and tried to adjust accordingly. It is possible that the adjustment over- or under-corrected the effect of wins. There are also a number of pitching and hitting metrics that are moderately to heavily correlated with wins, and these were not adjusted accordingly. Therefore, it is possible that some measure of confounding affected the outcome of this study.

Our labeling of teams as having a certain playing style assumes that playing style is uniform among all hitters or pitchers on a given team. While some teams do tend to prefer one skillset over another, teams do not have all identical pitchers or hitters.

We identified that a "flashy", Modern style of play, associated with high strikeouts, high stolen bases, and few walks seems to draw more fans both overall, and on the road. Future research could investigate if

this pattern has been apparent over the past decade in baseball, or whether it is apparent between leagues within MLB. We also concluded that the theoretical pace of play between teams does not appear to have an actual effect on attendance figures. This may be because fans do not perceive specific team as having a slow style of play, but rather the sport as a whole.

References

"2019 MLB Attendance - Major League Baseball - ESPN." ESPN.Com, http://www.espn.com/mlb/attendance/_/year/2019. Accessed 19 Mar. 2021.

"MLB Stats — Baseball Stats." MLB.Com, https://www.mlb.com/stats/team/2019. Accessed 19 Mar. 2021.

[7] Brownlee, Jason. "How to Scale Data With Outliers for Machine Learning." Machine Learning Mastery, 26 May 2020, https://machinelearningmastery.com/robust-scaler-transforms-for-machine-learning/.

[4] Butler, Michael R. "Interleague Play and Baseball Attendance." Journal of Sports Economics, vol. 3, no. 4, Nov. 2002, pp. 320–34.

[3] Davis, Michael C. "Analyzing the Relationship Between Team Success and MLB Attendance with GARCH Effects." Journal of Sports Economics, vol. 10, no. 1, Feb. 2009, pp. 44–58.

[5] Denaux, Zulal S., et al. "Factors Affecting Attendance of Major League Baseball: Revisited." Atlantic Economic Journal, vol. 39, no. 2, June 2011, pp. 117–27. Springer Link, doi:10.1007/s11293-011-9274-2.

[2] Johnson, Dave. "Overall Health for Professional Baseball in Trouble." WTOP, 10 June 2020, https://wtop.com/mlb/2020/06/overall-health-for-professional-baseball-in-trouble/.

Lohr, Tom. "How Major League Baseball Can Fix Its Attendance Problem." HowTheyPlay - Sports, https://howtheyplay.com/team-sports/How-Major-League-Baseball-Can-Fix-Its-Attendance-Problem. Accessed 19 Mar. 2021.

[6] MacDonald, Mark. "Does Bat Day We Cents? The Effect of Promotions on the Demand for Major League Baseball." Journal of Sport Management, vol. 14, no. 1, 2000, pp. 8–27.

[8] Soetewey, Antoine. "Outliers Detection in R." Stats and R, https://statsandr.com/blog/outliers-detection-in-
Accessed 19 Mar. 2021.

Woltring, Mitchell T. "Attendance Still Matters in MLB: The Relationship with Winning Percentage." The Sport Journal, 22 Nov. 2018, https://thesportjournal.org/article/attendance-still-matters-in-mlb-the-rel

[1] Zielonka, Adam, and David Driver. "Not so Fast: Baseball Vexed by Longer Games, Shrinking Attendance." The Washington Times, https://www.washingtontimes.com/news/2019/may/22/mlb-still-vexed-slower-game
Accessed 19 Mar. 2021.

[9] Albert, Jim. "Why Are Baseball Games So Long?" Exploring Baseball Data with R, 1 Jan. 2020, https://baseballwithr.wordpress.com/2020/01/01/why-are-baseball-games-so-long/.

References