MLB Attendance
Baseball’s Best Fan Bases
Michael Aronson, Jonathan Finch, Cody Justus, David Parker

What factors drive modern-day MLB attendance and what teams outperform attendance expectations?
Executive Summary

This study first and foremost seeks to determine what factors drive MLB attendance in the 21st century. Using these factors as a baseline, we then took an empirical approach in determining which fan bases outperform expectations and thus are inherently strong fan bases.

Motivation

- Game attendance, and associated ticket sales, is critically important in Major League baseball, a result of the 162 game season, far longer than any of the other major professional sports leagues
- Widespread debate on which teams have the “best” fan bases
- Numerous “rankings,” but typically qualitative in nature or based on a single season’s numbers
- Opportunity to fill this gap and take a data-driven approach to answering this age-old question

Past Research

- Relatively little research on the topic of fan loyalty at the team level
- Existing research has focused on the psychology behind and drivers of extreme fandom at the individual level
  1. Richard Kolbe and Jeffrey James determined that primary drivers are developing a passion for the team at a young age from the father and the feeling of a “community of fans.”
  2. Alan Tapp concurred regarding the importance of community and added that fans can become easily frustrated with a team and stop supporting them

Sources:
1. “An Identification and Examination of Influences That Shape the Creation of a Professional Team Fan” Kolbe, Richard H., James, Jeffrey D., International Journal of Sports Marketing & Sponsorship

Preview of Results

From our analysis, several factors revealed themselves to be significant including win percentage, team payroll, and the age of the team. Below are the MLB franchises that were determined to have the strongest and weakest fan bases:

<table>
<thead>
<tr>
<th>Strongest fan bases</th>
<th>Weakest fan bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Francisco Giants</td>
<td>Cleveland Indians</td>
</tr>
<tr>
<td>Los Angeles Angels</td>
<td>Washington Nationals</td>
</tr>
<tr>
<td>Milwaukee Brewers</td>
<td>Chicago White Sox</td>
</tr>
<tr>
<td>Boston Red Sox</td>
<td>Atlanta Braves</td>
</tr>
<tr>
<td>Houston Astros</td>
<td>Toronto Blue Jays</td>
</tr>
</tbody>
</table>
**METHODOLOGY**

**Variables of Interest [Source]**

*Dependent variable:* Attendance as % of capacity [ESPN]
- Control for differences in stadium size

*Independent Variables:*
1. Average Ticket Price (not including luxury suites) [Rodney Fort]
   - Natural log transformation to convert to percent changes
2. Payroll [USA Today]
   - Natural log transformation to convert to percent changes
3. Win% [baseballreference.com]
4. # of Pro Teams in Market [ESPN]
   - NFL, MLB, NBA, NHL, MLS
5. Founding Year* [mlb.com]
6. Time Since Last Championship [ESPN]
7. Playoffs?* [baseballreference.com]
8. American League?* [ESPN]

*Indicator Variable

**Summary Statistics**

**Attendance:**
Mean: 30,448 fans | Std. Dev.: 8,543
Range: 10,038 (Marlins ’02) - 53,069 (Yankees ’08)

**Ticket Price:**
Mean: $23.17 | Std. Dev.: 9.10
Range: $9.33 (Twins ’00) - $72.97 (Yankees ’09)

**Payroll:**
Mean: $81,796,445 | Std. Dev.: 37,337,589
Range: $15m (Marlins ’06) - $241m (Dodgers ’14)

**Win %:**
Mean: 50% | Std. Dec.: 7.09%
Range: 26.5% (Tigers ’03) - 71.6% (Mariners ’01)

**Number of Pro Teams in Market:**
Mean: 4.47 teams | Std. Dev.: 2.19
Range: 2 (Baltimore) - 10 (New York)

**Data Limitations**

- Time span: only covers 15 years, excluding a large portion of baseball’s long history
- Singular focus: only covers the MLB, excluding the other major professional sports leagues
- Variables: excludes certain (hard to measure) variables such as weather in a given region/city
**METHODOLOGY**

**Correlation Tests**

<table>
<thead>
<tr>
<th>Attendance as % of Capacity</th>
<th>Win %</th>
<th>ln(Ticket Price)</th>
<th>ln(Payroll)</th>
<th># of Pro Teams in Market</th>
<th>Founded: 1870-1900</th>
<th>Founded: 1901-1960</th>
<th>Founded: 1961-1990</th>
<th>Time Since Last Championship</th>
<th>Playoffs?</th>
<th>AL?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance as % of Capacity</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Win %</td>
<td>0.44</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(Ticket Price)</td>
<td>0.63</td>
<td>0.17</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ln(Payroll)</td>
<td>0.67</td>
<td>0.36</td>
<td>0.68</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of Pro Teams in Market</td>
<td>0.32</td>
<td>0.26</td>
<td>0.35</td>
<td>0.44</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Founded: 1870-1900</td>
<td>0.28</td>
<td>0.12</td>
<td>0.07</td>
<td>0.16</td>
<td>-0.02</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Founded: 1901-1960</td>
<td>0.07</td>
<td>0.14</td>
<td>0.28</td>
<td>0.16</td>
<td>0.18</td>
<td>-0.36</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Founded: 1961-1990</td>
<td>-0.14</td>
<td>-0.15</td>
<td>-0.13</td>
<td>-0.10</td>
<td>-0.09</td>
<td>-0.43</td>
<td>-0.43</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Since Last Championship</td>
<td>0.09</td>
<td>-0.14</td>
<td>0.12</td>
<td>-0.07</td>
<td>-0.07</td>
<td>0.08</td>
<td>-0.05</td>
<td>0.18</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Playoffs?</td>
<td>0.34</td>
<td>0.70</td>
<td>0.14</td>
<td>0.27</td>
<td>0.18</td>
<td>0.13</td>
<td>0.14</td>
<td>-0.20</td>
<td>-0.16</td>
<td>1.00</td>
</tr>
<tr>
<td>AL?</td>
<td>-0.11</td>
<td>0.05</td>
<td>0.10</td>
<td>0.05</td>
<td>0.06</td>
<td>-0.57</td>
<td>0.64</td>
<td>0.06</td>
<td>-0.10</td>
<td>0.03</td>
</tr>
</tbody>
</table>

**Results/Implications**

- Ran correlation tests in order to test for potential issues of collinearity
- Payroll and ticket price were shown to be partially correlated, but due to their importance to the interpretation of our model (strongly correlated with attendance), both variables were kept, and ultimately both proved to be statistically significant
- AL indicator was removed due to collinearity with the year founded variables
Modeling Technique: Multiple Linear Regression

Initial Regression Results:

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.5714</td>
<td>0.2881</td>
</tr>
<tr>
<td>Win %</td>
<td>0.5948*</td>
<td>0.1134</td>
</tr>
<tr>
<td>Ln(Ticket Price)</td>
<td>0.1916*</td>
<td>0.0237</td>
</tr>
<tr>
<td>Ln(Payroll)</td>
<td>0.1284*</td>
<td>0.0188</td>
</tr>
<tr>
<td># of Pro Teams in Market</td>
<td>0.0011</td>
<td>0.0029</td>
</tr>
<tr>
<td>Founded: 1870-1900</td>
<td>0.0613*</td>
<td>0.0205</td>
</tr>
<tr>
<td>Founded: 1901-1960</td>
<td>-0.0254</td>
<td>0.0204</td>
</tr>
<tr>
<td>Founded: 1961-1990</td>
<td>0.0018</td>
<td>0.0193</td>
</tr>
<tr>
<td>Time Since Last Championship</td>
<td>0.0008*</td>
<td>0.0003</td>
</tr>
<tr>
<td>Playoffs?</td>
<td>0.0204</td>
<td>0.0174</td>
</tr>
</tbody>
</table>

R^2 0.6041

Observations 450

Key Results:

- Statistically significant and strong positive coefficients for Win% and Ln(Payroll)
- Significant and positive coefficient for Founded: 1870-1900
- (Surprisingly) positive coefficient for Ln(Ticket Price)

Implications:

- Improvements in on-field performance and increases in payroll, which can be viewed as a proxy for ownership’s commitment to the team’s success, significantly increase attendance
- Baseball’s most historic teams, those founded between 1870-1900, have an inherent attendance advantage
- Lowering ticket price is not an effective means of increasing attendance and is unlikely to compensate for poor on-field performance
Empirical Approach

The above model predicts a team’s attendance based on all of the listed factors. The error term ($\varepsilon$) thus captures any other variables that could explain attendance but are not included in the model. In our effort to include the majority of the measurable variables that could drive attendance, we believe the error term would effectively capture the unmeasurable effect of the inherent strength of fan bases. Under this belief, the difference between our prediction and the actual attendance would provide a metric for measuring this inherent factor. The results of this empirical approach are shown on the following slide.

Iterative Process

**Step 1: Correlation Analysis**
[Check for collinearity]
- Removed American League variable (high correlation with Year Founded variables)

**Step 2: Initial Regression**
[Check for statistical significance]
- Removed # of Pro Teams in Market variable
- Removed Founded (1901-1960 & 1961-1990) variables
- Removed Playoffs variable

**Steps 3 & 4: Final Regression & Predictive Model**
- With variables removed, ran a final regression
- The coefficients from this regression were used to build our predictive model

**Predictive Model**

\[
\text{Attendance} = -2.6252 + 0.6623\times\text{Win}\% + 0.1829\times\ln(\text{Ticket Price}) + 0.1311\times\ln(\text{Payroll}) + 0.0711\times\text{Founded:1870-1900} + 0.0008\times\text{Time Since Last Championship} + \varepsilon
\]
**Insights**

- Many teams with positive differentials are consistent with common perceptions of teams with strong fan bases (Giants, Red Sox, Cardinals, Tigers, and Cubs).
- Some of the more surprising teams (Brewers, Astros, Twins, Pirates) most likely show up in our list for two reasons:
  - (1) Teams with smaller stadiums perform better in our study, considering our dependent variable is percentage of stadium capacity filled.
  - (2) Our model predicts that these teams would have weaker attendance due to their performance with respect to our independent variables (win percentage, payroll, etc.), but they have been able to exceed the notably low expectations.
- Similarly, several teams in the bottom half of the rankings stand out as surprises, namely the Dodgers and Yankees:
  - These teams are subject to the flip side consequences of reasons (1) and (2) above. They have well above average size stadiums and thus are at a disadvantage within our model with respect to percentage of capacity filled instead of total attendance.
  - Furthermore, based on these teams’ consistently strong performance (i.e. playoffs, win percentage, payroll), our model predicts extremely high attendance numbers that might be unachievable in the real world.
- **In sum, we believe our model serves as a credible basis to examine the relative strengths of fan bases for MLB teams.**

**Potential Extensions**

- Extend list of independent variables to try to capture additional explanatory power for some of the more surprising findings (new stadium, median income, market size, etc.)
- Expand to other pro sports and look for trends within cities.