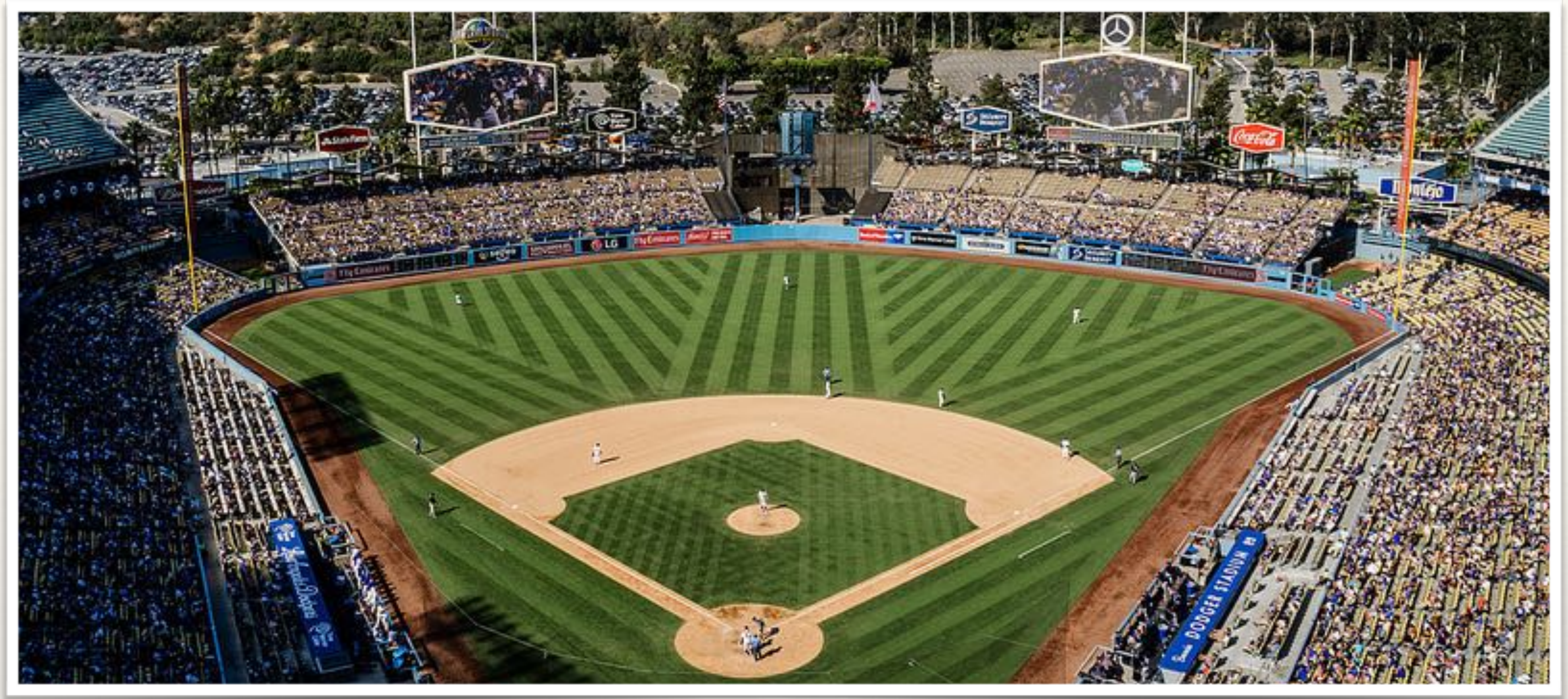


# Major League Baseball (MLB) Attendance

Leveraging data science to create a critical product for MLB Organizations

# Background & Context

- MLB features 30 teams — 29 in the United States across all major cities, and 1 in Toronto, Canada.

- MLB consist of 2 leagues — American and National with 3 divisions each

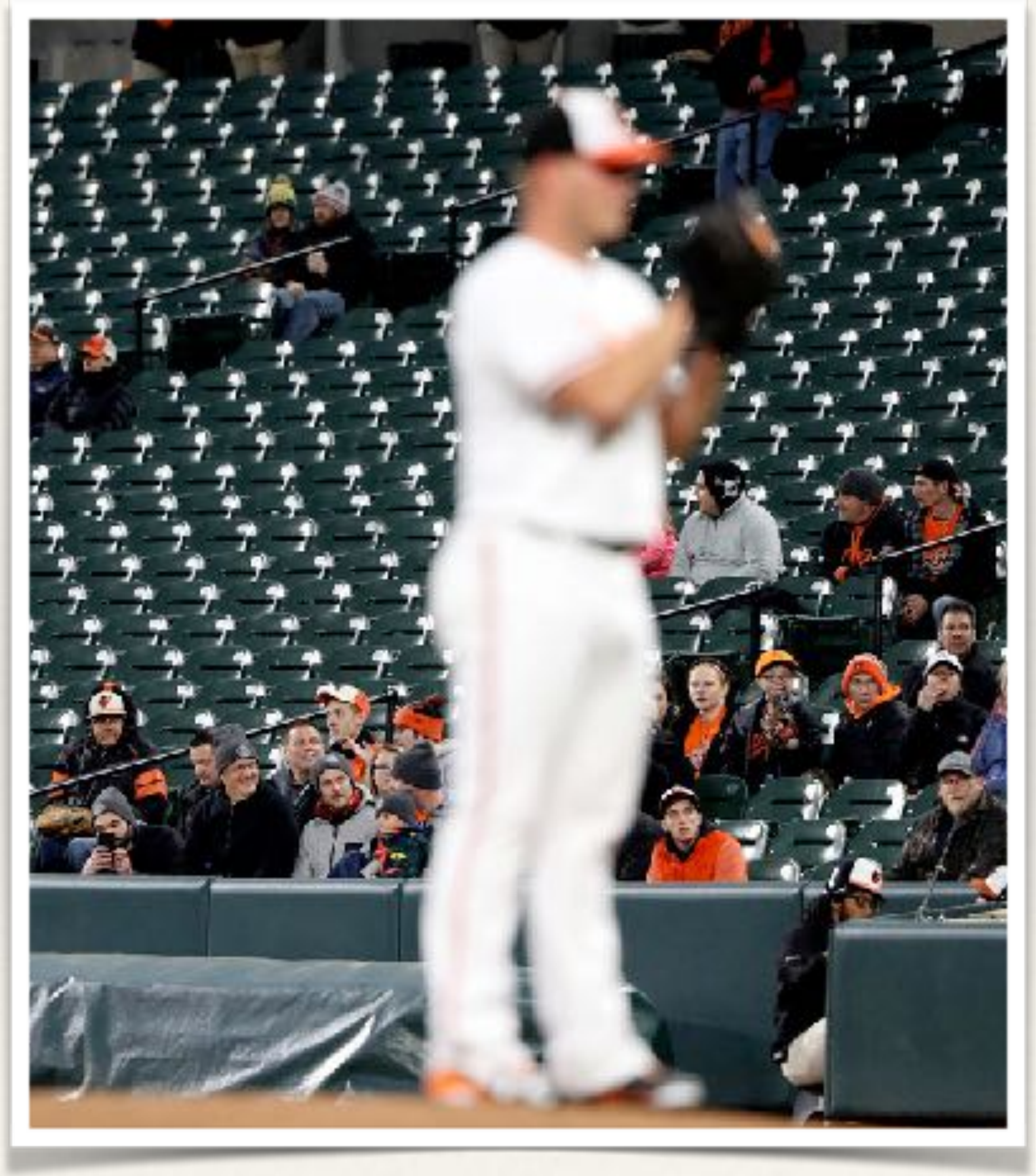
- Eclipsed $10B in total annual revenue in 2017, behind only the National Football League (NFL).

- First World Series was held in 1903. Baseball has been a mainstay throughout American history and is considered by many to be our “national pastime”.

# Current Trends

- Overall MLB attendance for 2017 represents the third consecutive season of decline and the lowest mark since 2002.

- Viewership average age increased from 52 to 57 between 2000-2016. Other major sports (NFL & NBA) retain a lower age and rate of increase in the same timeframe.



# Key Question

How do teams supply their pricing departments (ticketing, labor, supplier relations, concessions) with recommendations for revenue optimization?

# Implementation

- Used data from baseball-reference.com:
  - Source game record data for each team for both the 2016 and 2017 seasons
  - Total of 60 individual pages sources
- Used data from vividseats.com:
  - Source average ticket price per team for 2016 and 2017
- Manually found stadium capacity for each team

# Features

| Raw Data | Engineered Features |
|---|---|
| Date | Attendance to Capacity Ratio |
| Team | Attendance Category |
| Attendance | Early Season Indicator |
| Home or Away | Mid-Season Indicator |
| Opponent | Post-Season Indicator |
| Win/Loss | Hot Streak (3+ consecutive wins) |
| Runs Scored | Average Revenue |
| Runs Against | Day |
| Division Rank | Month |
| Games Back in Division | Year |
| Winning Pitcher | Weekday/Weekend |
| Losing Pitcher | |
| Time of Game | |
| Day/Night | |
| Streak | |
| Win Percentage | |
| Average Ticket Price | |
| Stadium Capacity | |

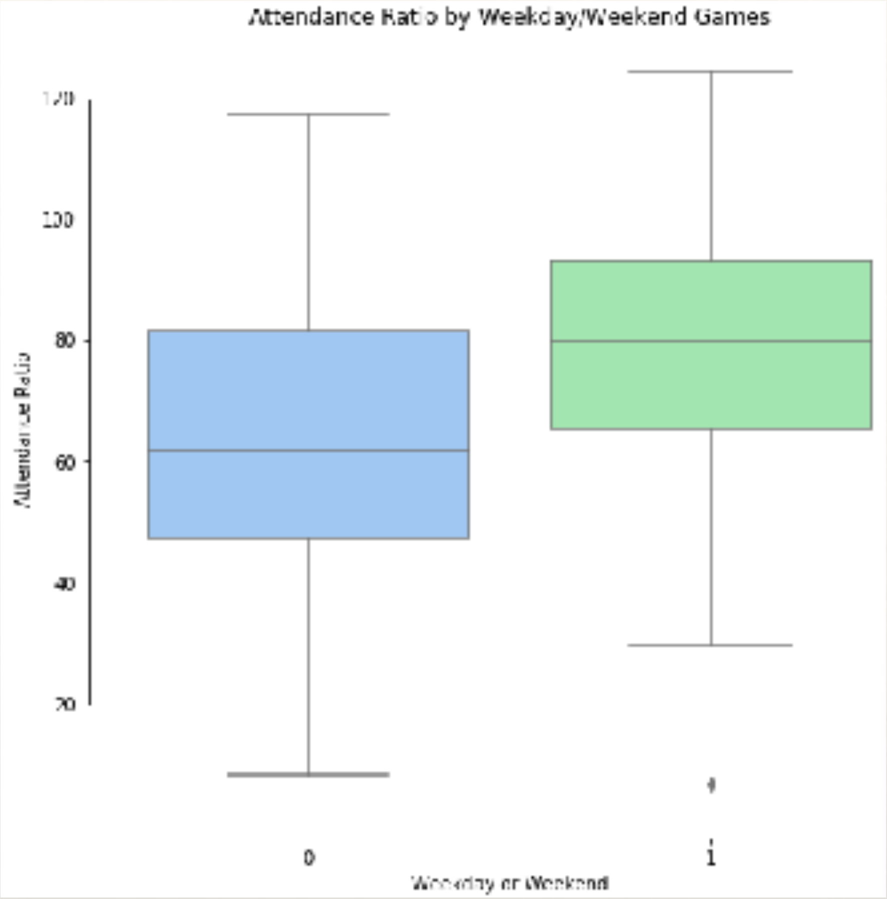
- All raw columns required a significant amount of cleaning/reengineering from web scraping
    - Splitting/stripping strings
    - Type conversion
    - Binary conversion
    - Label encoders (categoricals)
    - Functions to extract compatible data from ‘Games Back in Division’, ‘Time of Game, and ‘Streak’

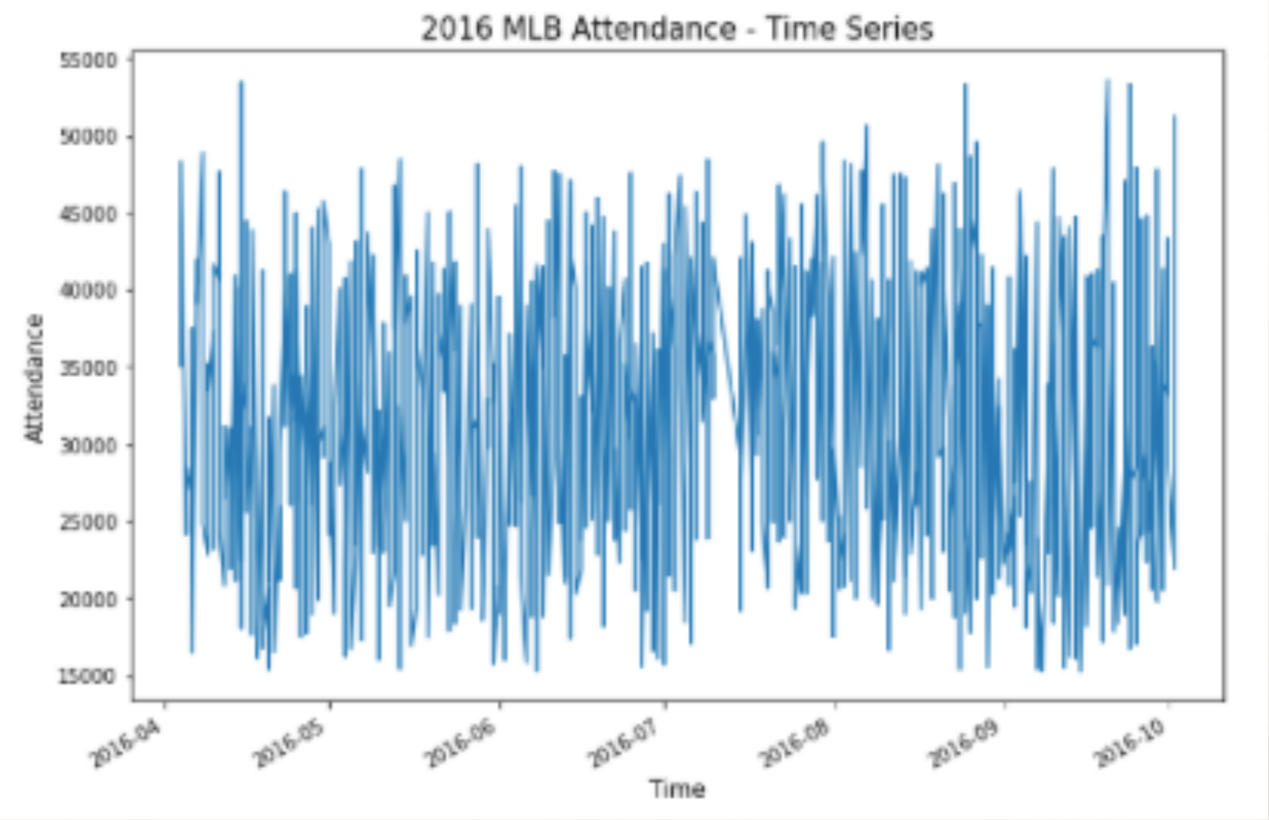
# Proposed Product - Recommendation Engine

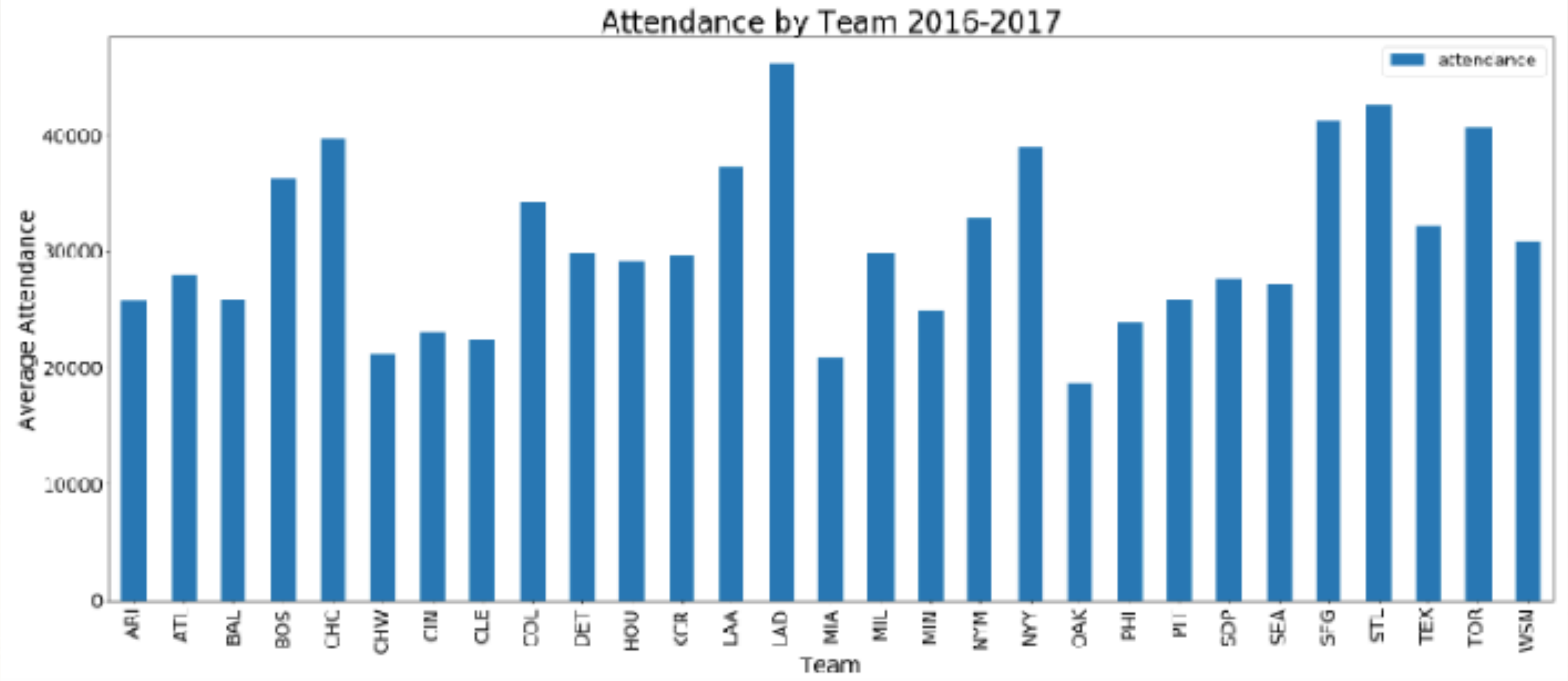
Recommendation Engine Components

| Component | Priority | Measure | Benefits |
|---|---|---|---|
| Attendance Category | Primary | • Capacity Less than 50%<br>• Capacity Between 50 and 75%<br>• Capacity Between 75 and 100%<br>• Capacity Greater than 100% | • Ticket prices could be more strategically established based on expected attendance category (based on stadium section)<br>• Savings on labor and supplier relations as only what is required can be planned based on how many people are expected |
| Weekday vs. Weekend Game | Secondary | Weekday — Mon, Tues, Wed, Thurs<br>Weekend — Fri, Sat, Sun | • Promotions and special events could be better planned<br>• Further supports ticketing department |
| Day vs. Night Game | Secondary | Day — afternoon start time<br>Night — evening start time | • Promotions and special events could be better planned<br>• Further supports ticketing department |
| Time Series Analysis | Tertiary | AIC, P-values, Log Likelihood | • Being able to predict actual attendance value for a game within a certain confidence interval would allow for an even more precise measure than an attendance category |

# Quick Glimpse Into Exploration







# Methodology

- Supervised Learning Techniques (Classification)
  - KNN, Random Forest, Gradient Boosting, Logistic Regression, Ridge Logistic Regression
- Unsupervised Techniques (Clustering)
  - KMeans
- Time Series Techniques
  - PACF, ARIMA Modeling

# Results

## Supervised Learning (Variety of Techniques)

| Prediction | Baseline | Best Model | Training Accuracy | Test Accuracy (Cross-Val) | Baseline vs. Test Difference | Observations |
|---|---|---|---|---|---|---|
| Attendance Category | 42% | Ridge Logistic Regression | 73% | 71% | Increase of 29% | Slight overfitting; Ran in 1.27 seconds |
| Weekday/Weekend | 50% | Random Forest | 100% | 65% | Increase of 15% | Slight overfitting; Ran in 7.95 seconds |
| Day/Night | 66% | Random Forest | 100% | 80% | Increase of 14% | Slight overfitting; Ran in 6.16 seconds |

## Unsupervised Learning (KMeans)

| Prediction | # Clusters | Clusters Representative (Y/N) | Observations |
|---|---|---|---|
| Attendance Category | 3 | Yes | Elbow curve suggested 3 clusters, as opposed to the 4 attendance categories. Hard time identifying > 100% |
| Weekday/Weekend | 3 | No | Not enough data to accurately cluster |
| Day/Night | 3 | No | Not enough data to accurately cluster |

## Time Series Analysis (PACF, ARIMA)

| Prediction | PACF Outliers | ARIMA Order | AIC | Log Likelihood | P-Value | Observations |
|---|---|---|---|---|---|---|
| 2016 Season | Minimal | 1, 1, 1 | 21005.985 | -10498.992 | Const 0.645<br>AR 0.451<br>MA 0.00 | N/A |
| 2017 Season | Minimal | 0, 1, 1 | 21000 | -10497.247 | Const 0.838<br>MA 0.00 | N/A |

# Future Work & Limitations

- Additional data sources
    - Team by team, game by game, ticket price information (down to the section)
    - Weather-related data
    - 10-20 years worth of attendance/game record data for each team
    - Find other ways to measure games where attendance is greater than capacity
- Pilot with one team for upcoming season
    - See how this information enables pricing departments and compare revenue statistics to previous season
    - Use parametric/non-parametric tests to see if populations are truly different

# References

- https://www.forbes.com/sites/maurybrown/2017/11/22/mlb-sets-record-for-revenues-in-2017-increasing-more-than-500-million-since-2015/#382b8dd37880
- https://www.history.com/this-day-in-history/national-league-of-baseball-is-founded
- https://www.forbes.com/sites/maurybrown/2017/10/02/final-2017-mlb-attendance-dips-below-73-million-for-first-time-since-2002/#4e350469326f
- https://www.marketwatch.com/story/the-sports-with-the-oldest-and-youngest-tv-audiences-2017-06-30