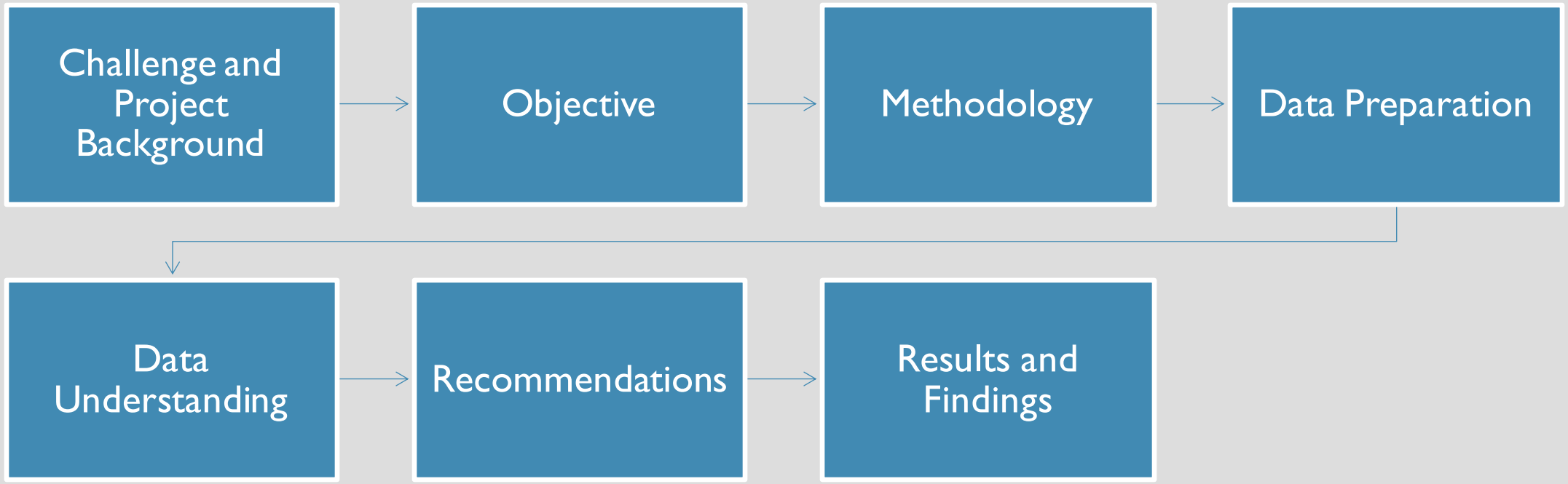
BAIS CAPSTONE PROJECT

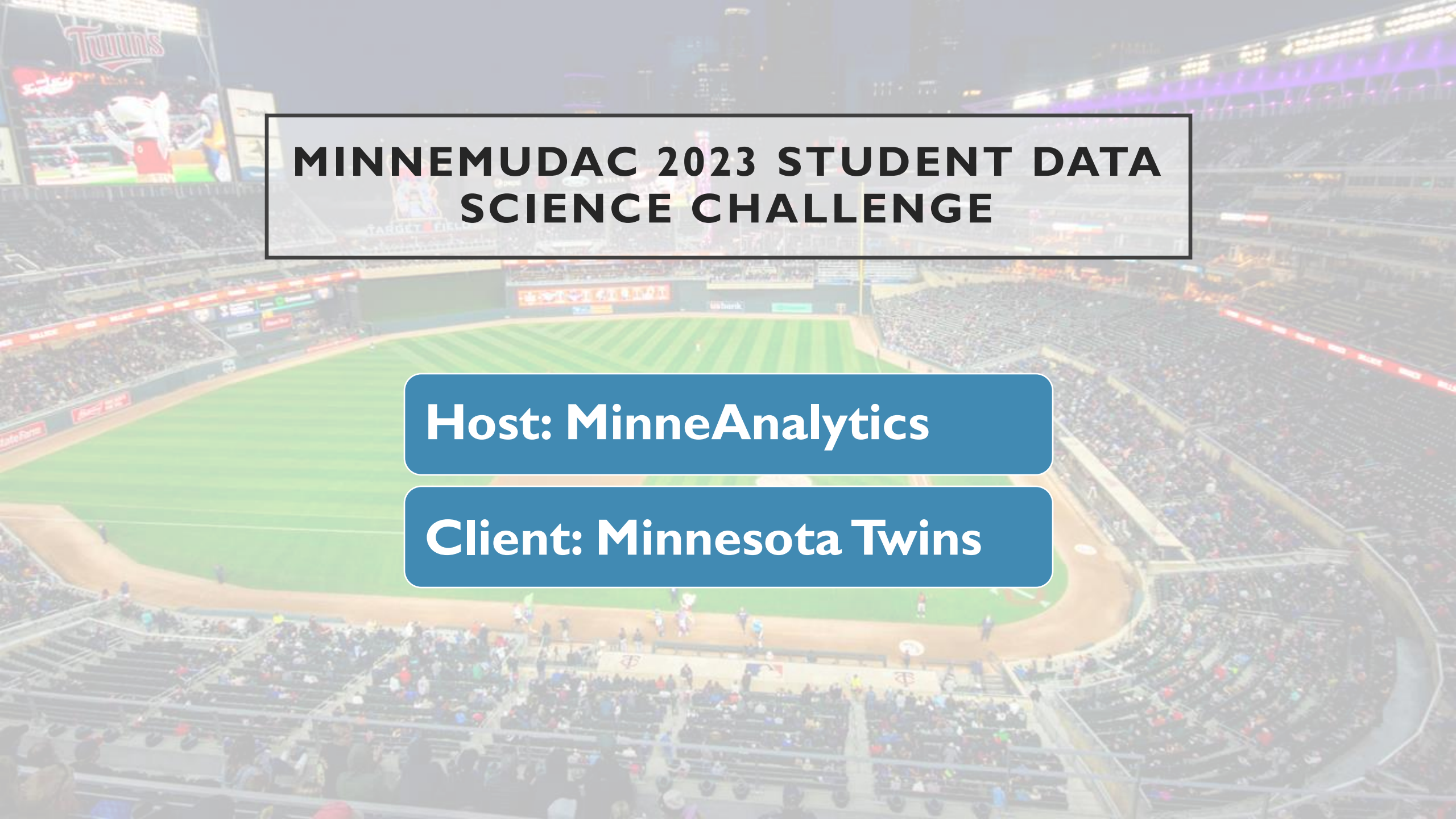
Home Attendance for the Minnesota Twins

MLB Group

Beth Tiggelaar, Katie Zawoyski, Ryan Yuson, Sarah Phillips

AGENDA


Challenge and Project Background
Objective
Methodology
Data Preparation
Data Understanding
Recommendations
Results and Findings

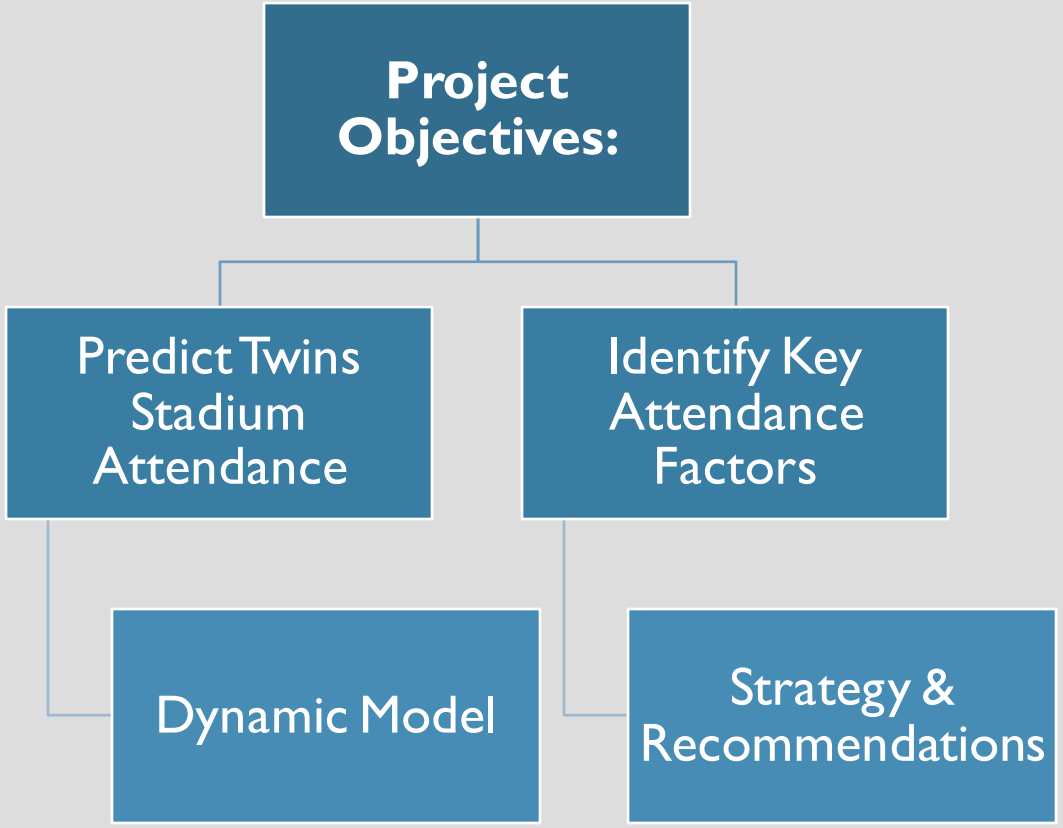


MINNEMUDAC 2023 STUDENT DATA SCIENCE CHALLENGE

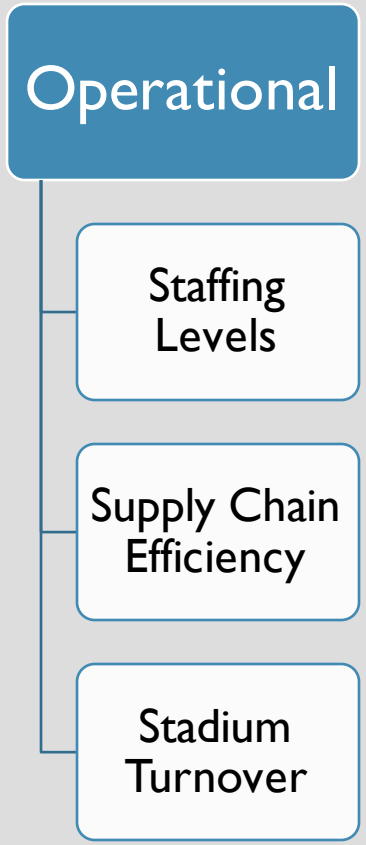
Host: MinneAnalytics

Client: Minnesota Twins

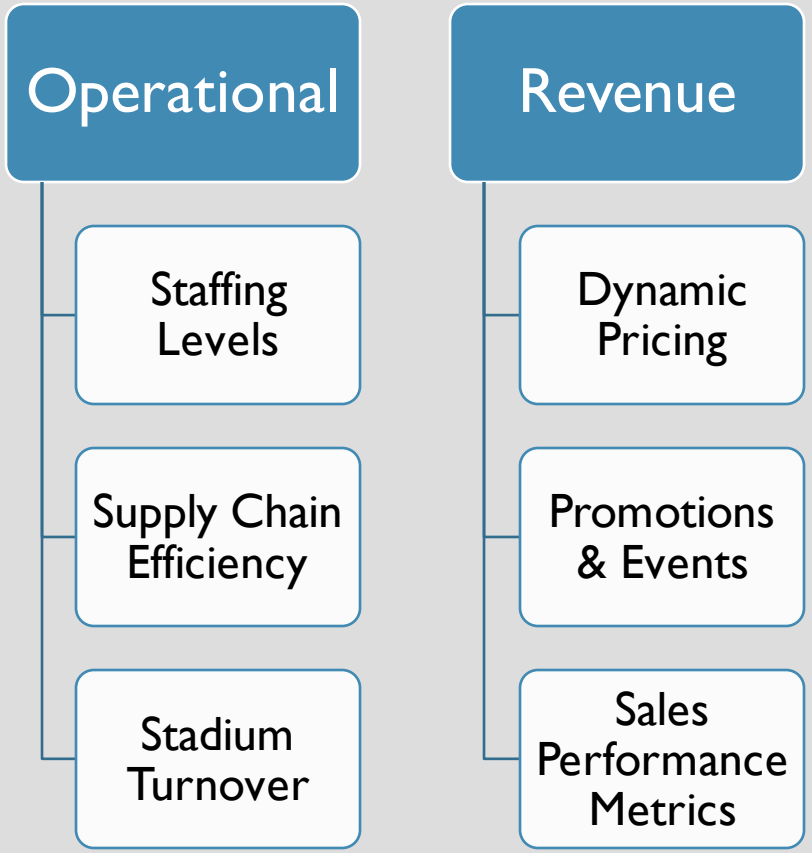
UNDERSTAND MLB HOME GAME FACTORS TO INFORM ATTENDANCE GROWTH



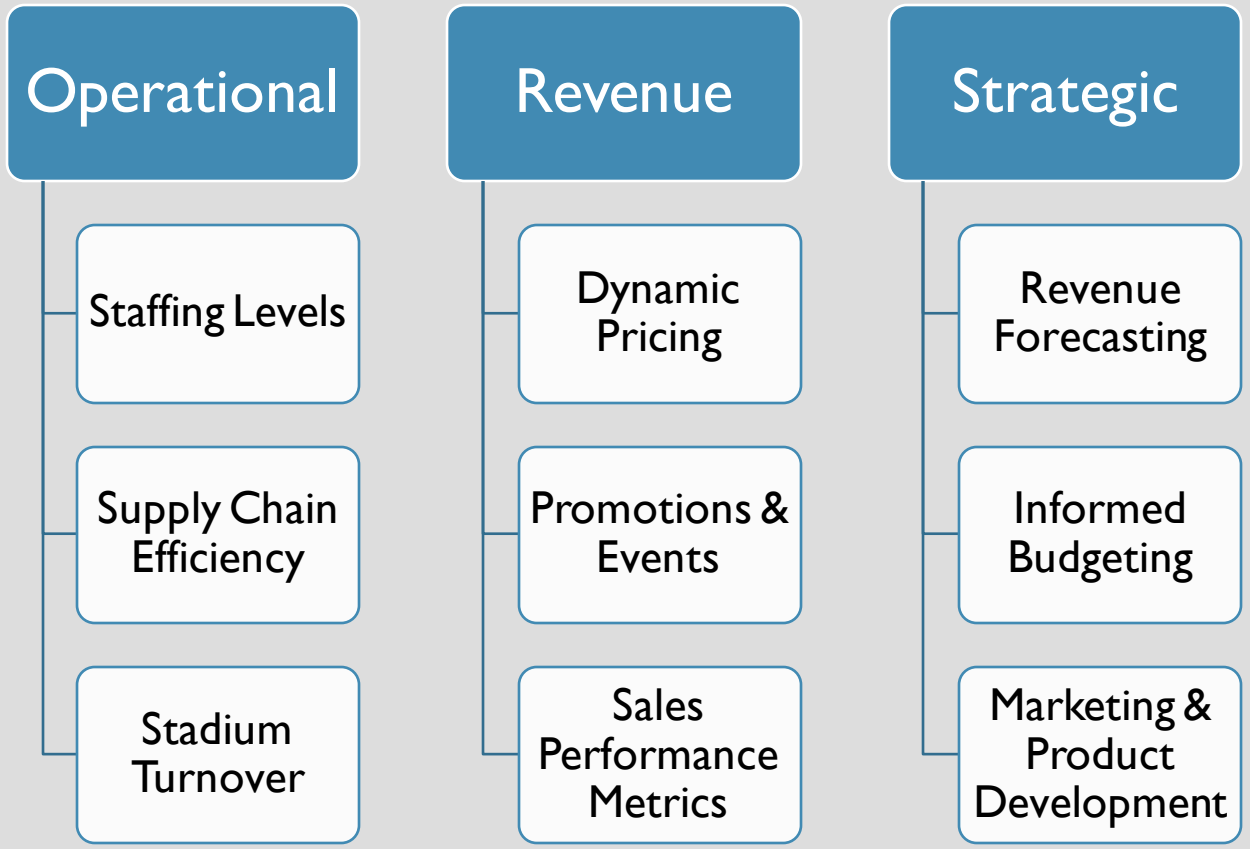
PREDICTING ATTENDANCE FOR BUSINESS PROCESS OPTIMIZATION

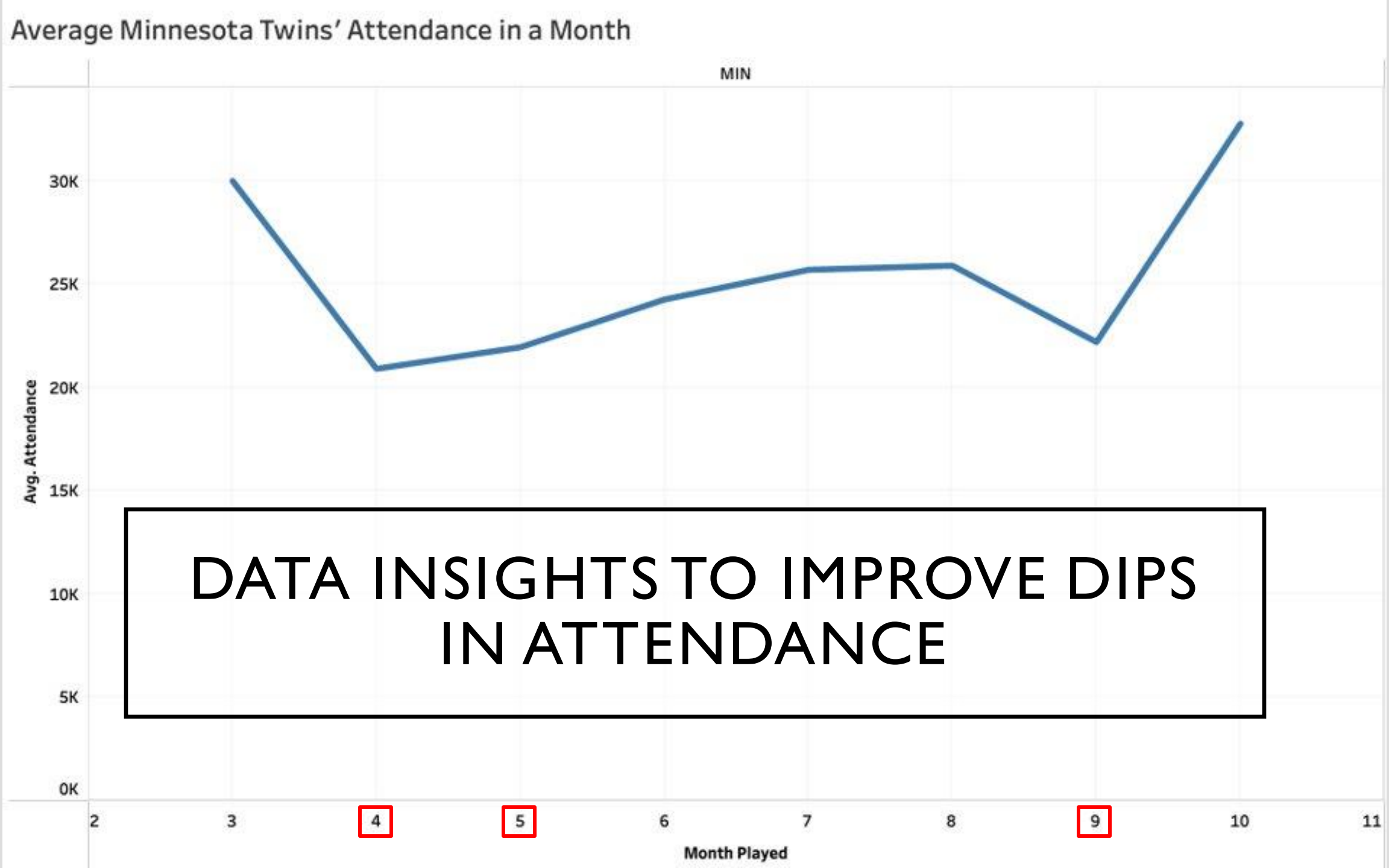


PREDICTING ATTENDANCE FOR BUSINESS PROCESS OPTIMIZATION



PREDICTING ATTENDANCE FOR BUSINESS PROCESS OPTIMIZATION





DATA INSIGHTS TO IMPROVE DIPS IN ATTENDANCE

METHODOLOGY

T-test & Variable Importance

METHODOLOGY

T-test & Variable Importance

Initial Regression Model

METHODOLOGY

T-test & Variable Importance

Initial Regression Model

Collect External Data

METHODOLOGY

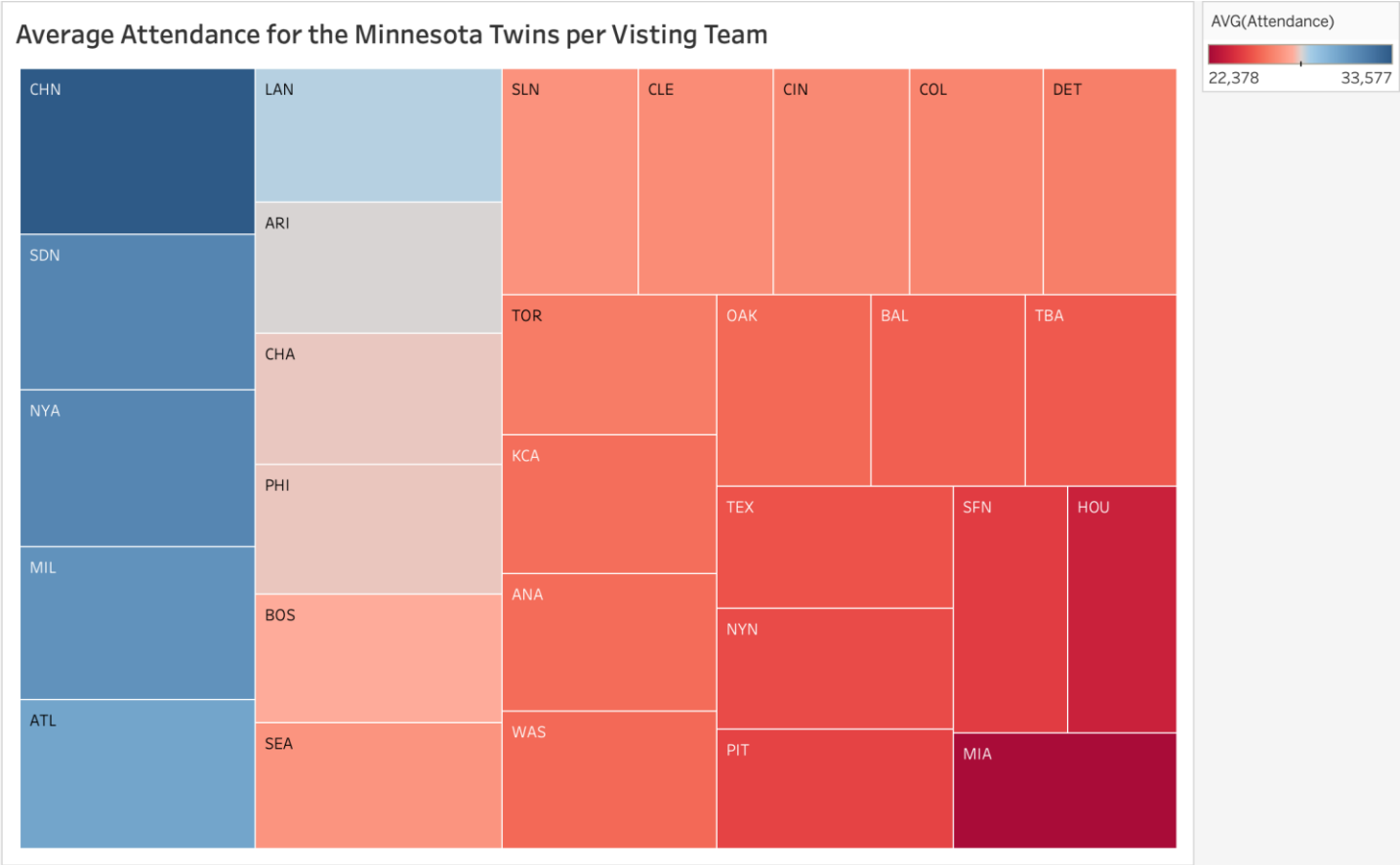
- T-test & Variable Importance
- Initial Regression Model
- Collect External Data
- Integrate into Final Regression

DATA PREPARATION

1. Properly format data types
2. Create dummy variables from categoric features
3. Handle column position for consistency
4. Remove instances with invalid attendance
5. Gather and integrate external data
6. Create subsets from GameLogs

AREAS OF DATA EXPLORATION

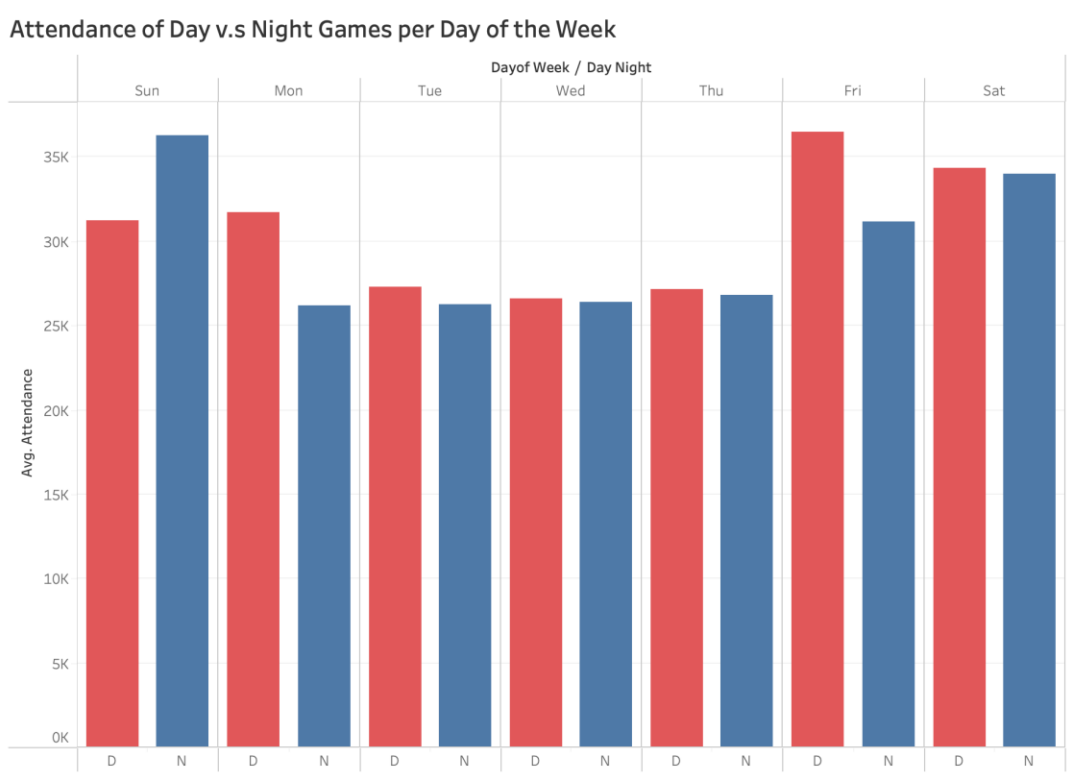
- Minnesota Twin's Average Attendance
- MLB Attendance Time Comparisons
- League Attendance Comparisons
- Variable Importance
- Attendance by Location

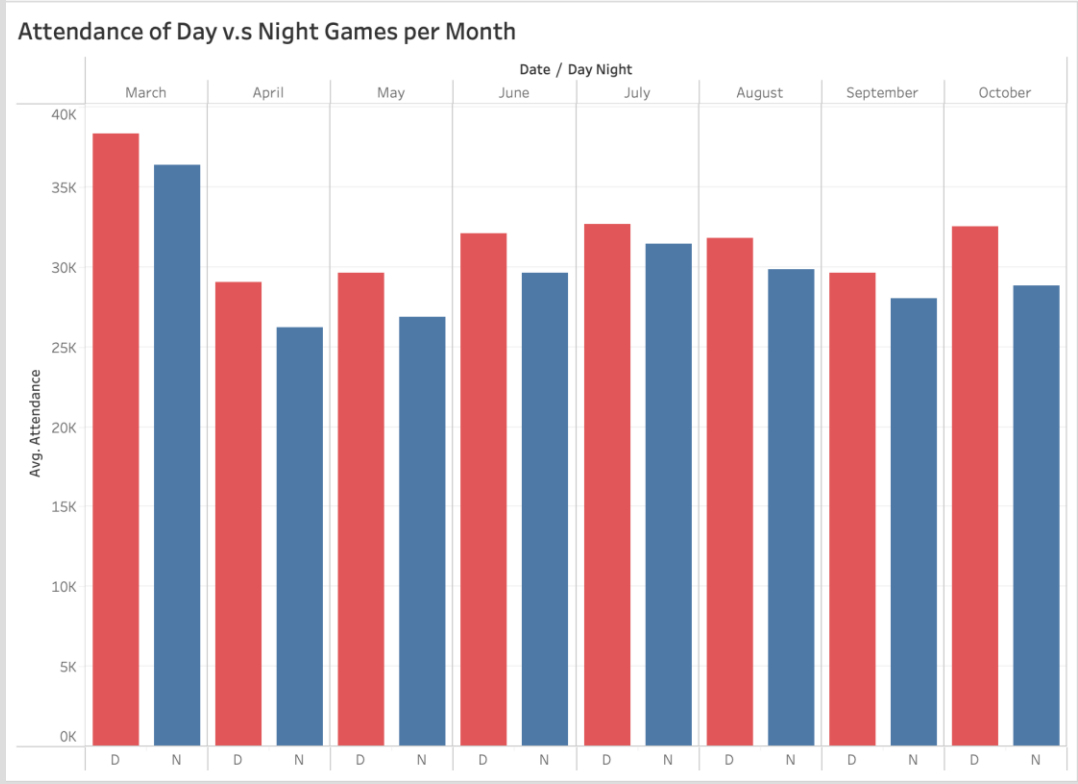


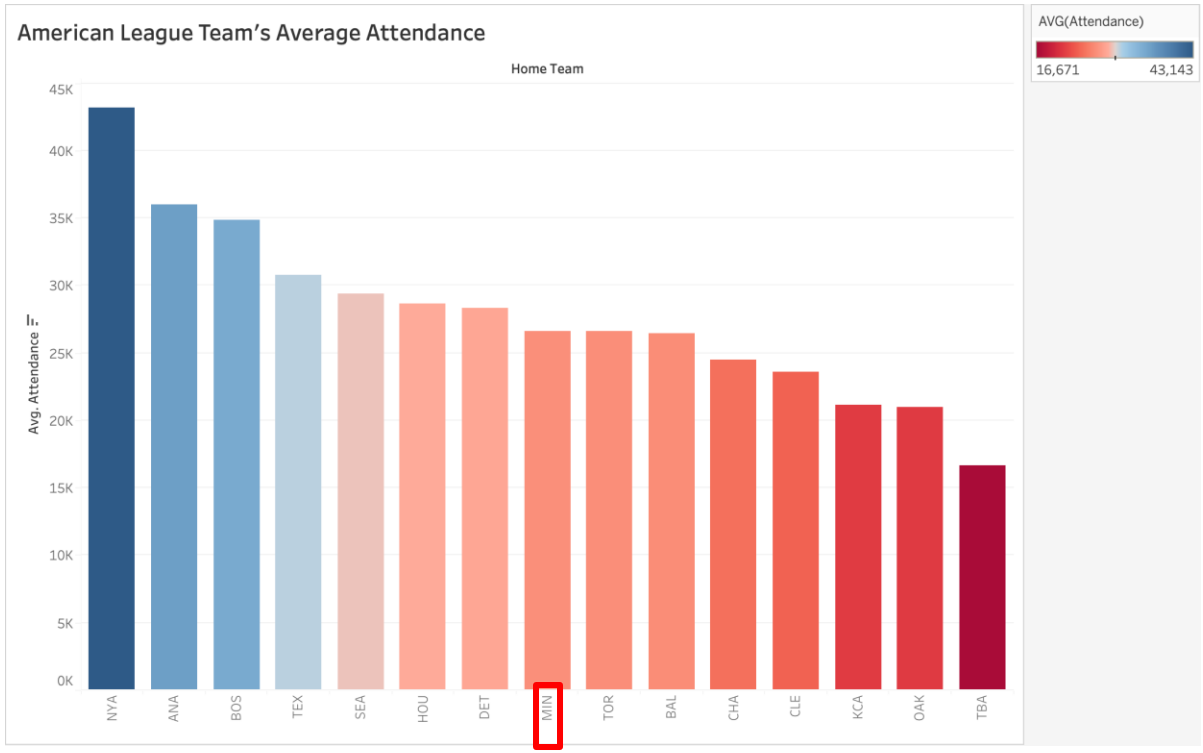
Chicago Cubs have the **highest** attended games on average
Miami Marlins have the **lowest** attended games on average

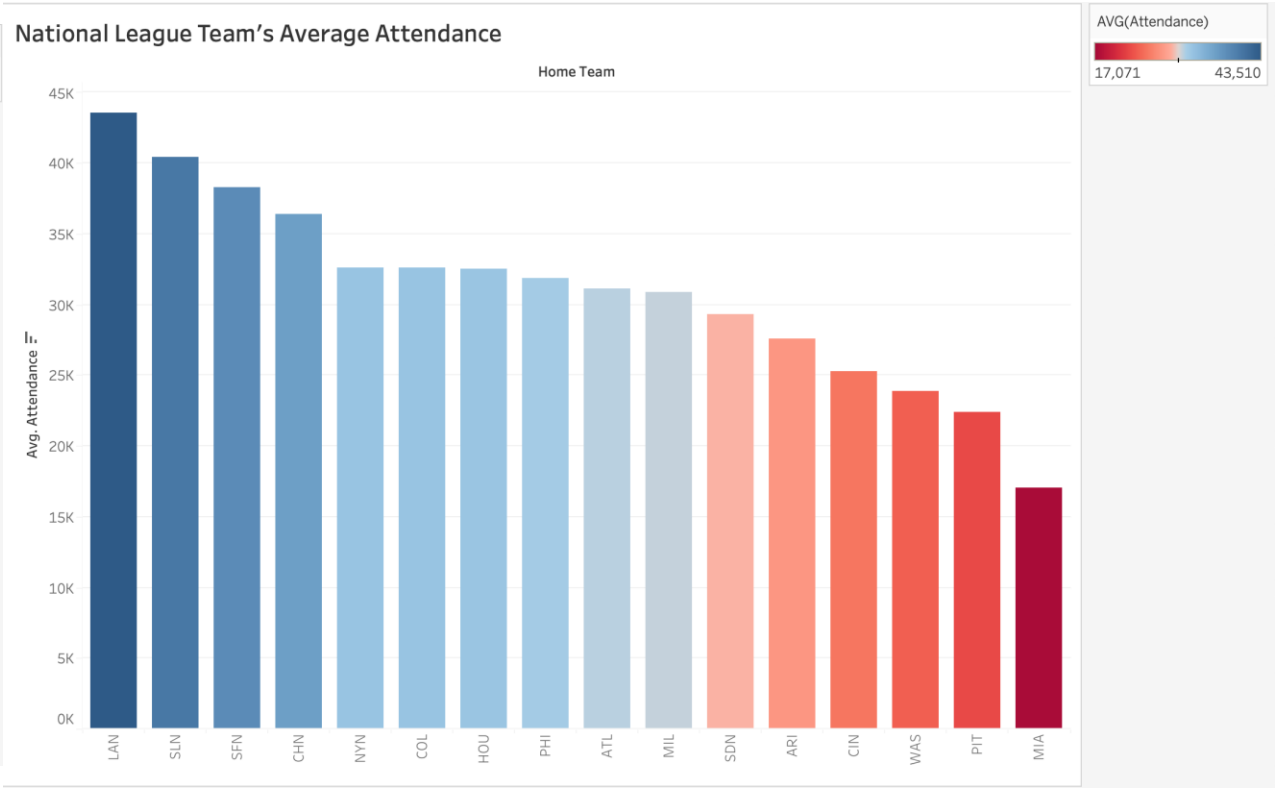
MINNESOTA TWINS ATTENDANCE COMPARED TO VISITING TEAM ATTENDANCE

DAY GAMES AVERAGE **HIGHER** ATTENDANCE THROUGHOUT MLB



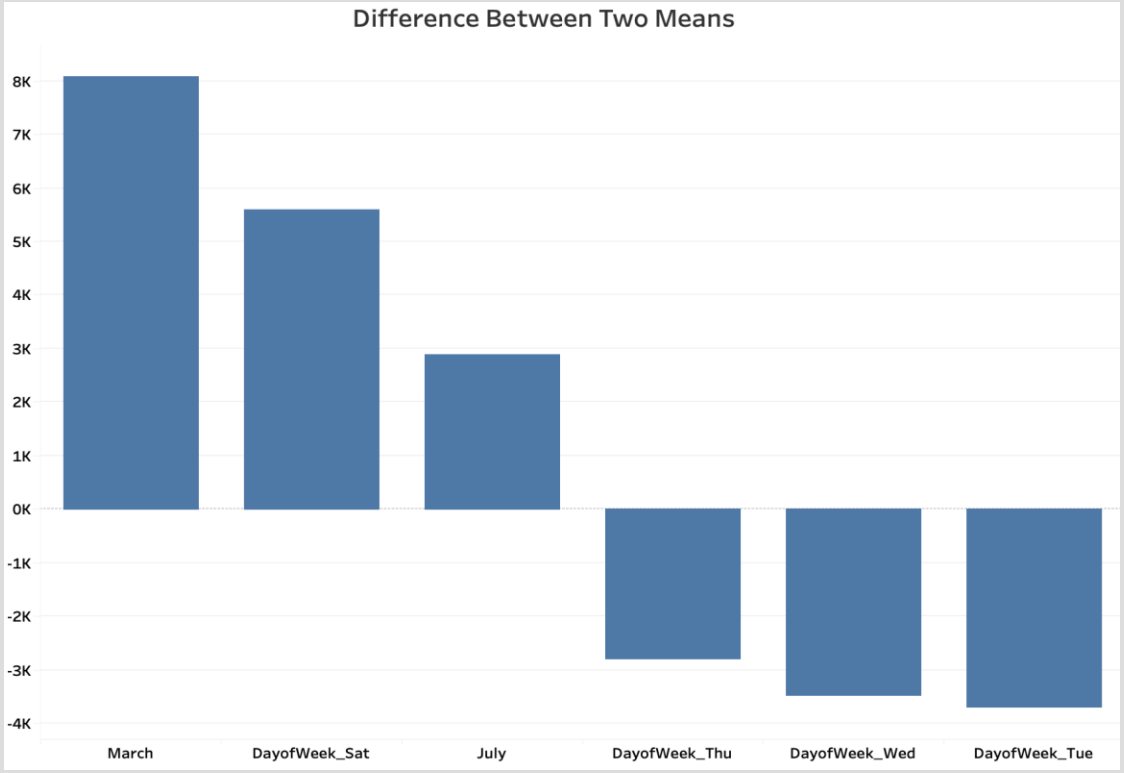




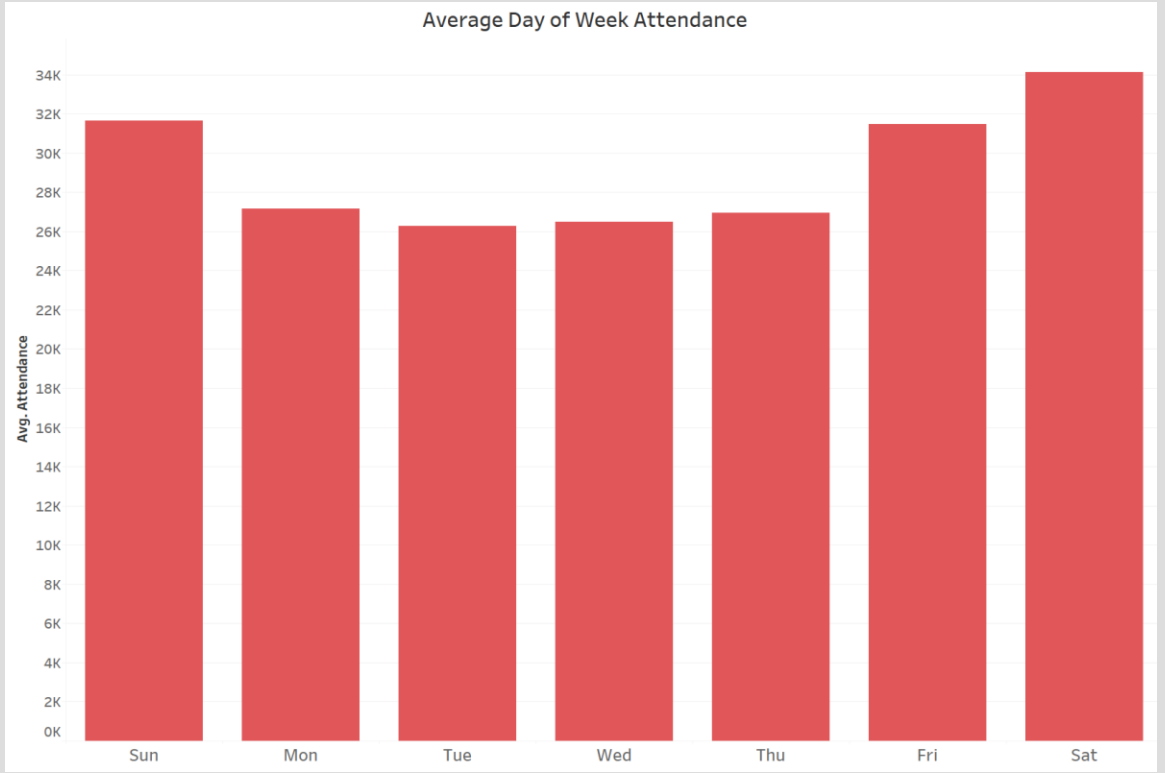


ATTENDANCE VARIANCE BY LEAGUE

VARIABLE IMPORTANCE PROVIDES DATA INSIGHTS



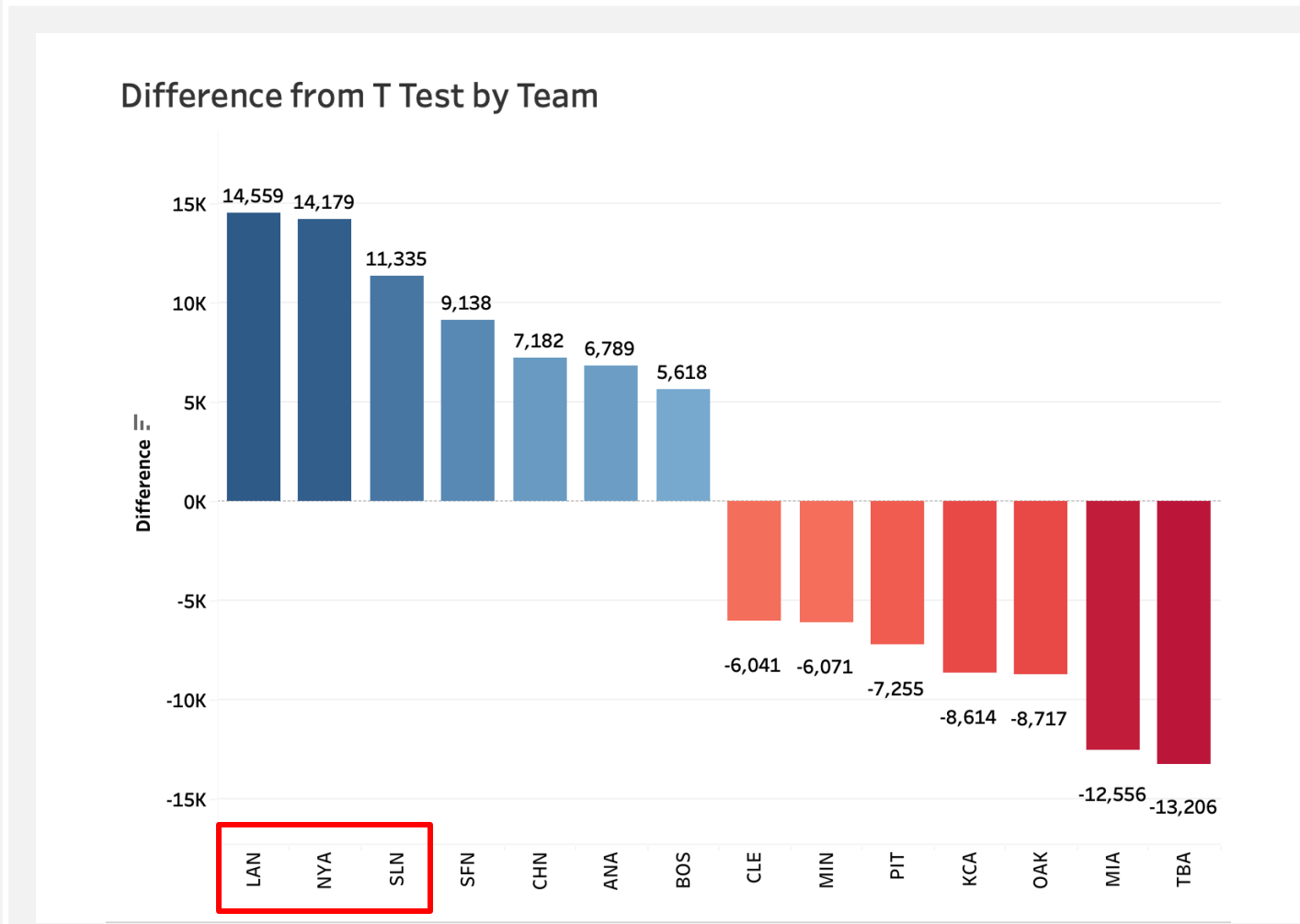
March, July, and **Saturday** games have a **positive impact** on attendance



Weekdays average the **lowest** Attendance

HIGHLY IMPACTFUL TEAMS HELP INFORM BUSINESS STRATEGY

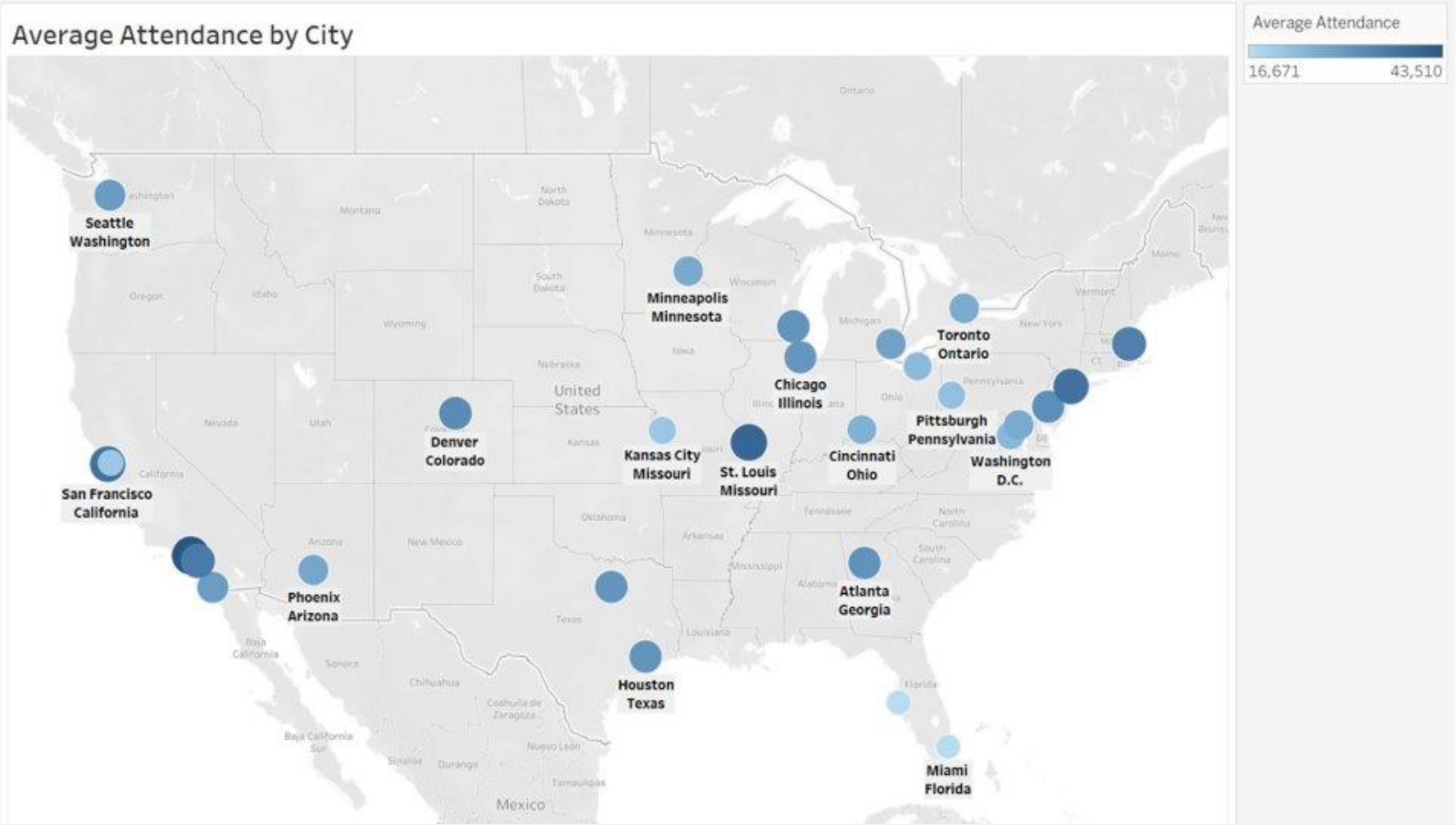
- **Los Angeles Dodgers, New York Yankees,** and **St. Louis Cardinals** have largest positive impact on attendance



EXTERNAL DATA PROVIDES A HOLISTIC APPROACH TO STADIUM INSIGHTS

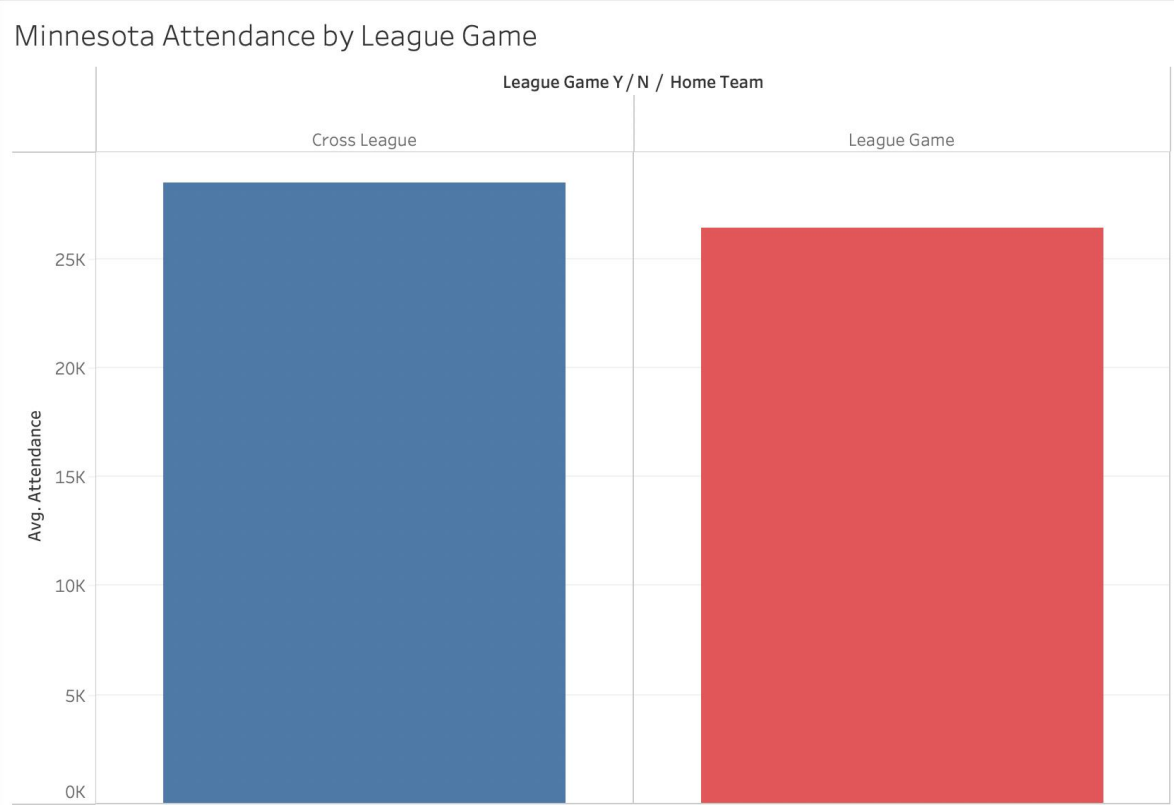
Modeling

STADIUM LOCATION AFFECTS AVERAGE ATTENDANCE

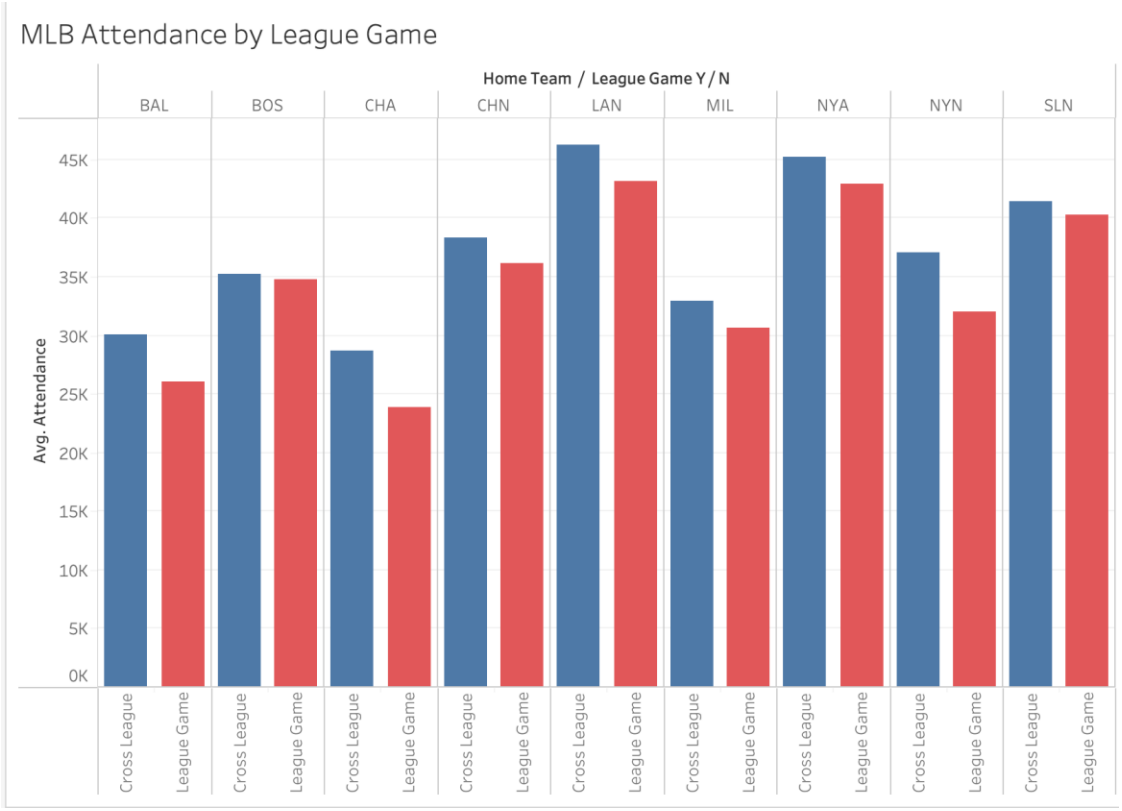


- **Los Angeles, CA** and **St. Louis, MO** have the highest average attendance
- **St. Petersburg, FL** and **Miami, FL** have the lowest average attendance followed by **Oakland, CA**

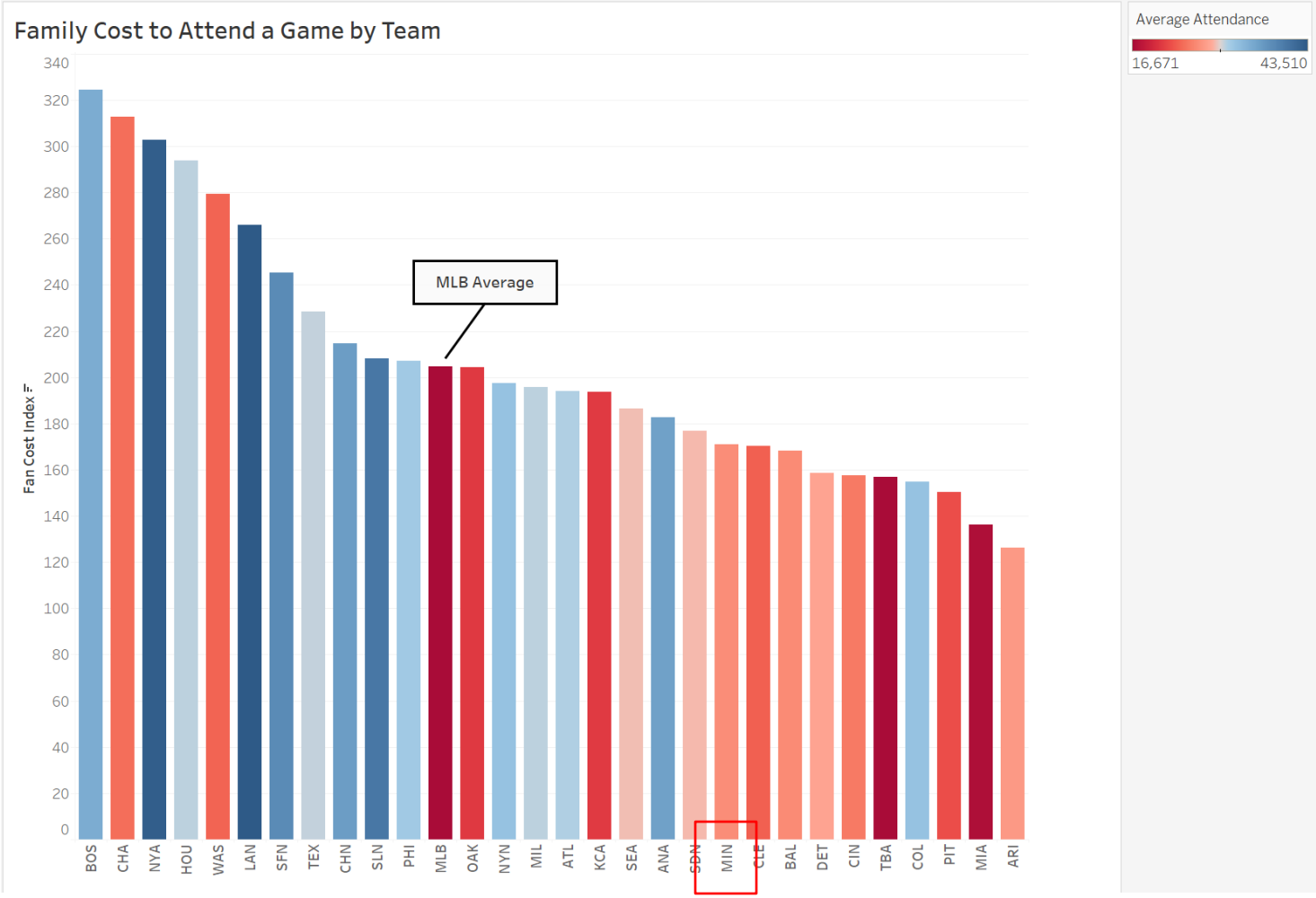
Minnesota Average Attendance by League



Other MLB Average Attendance by League



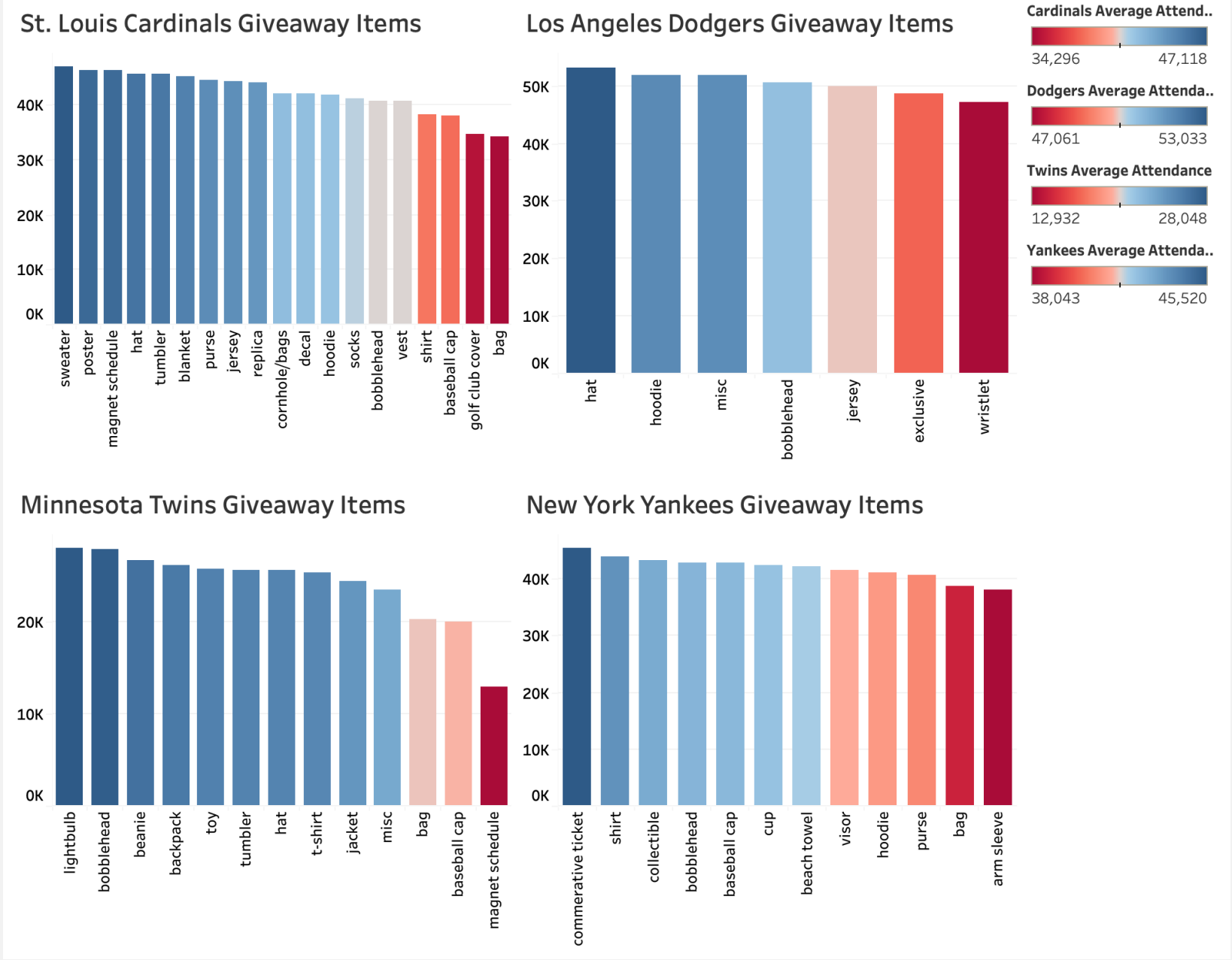
CROSS LEAGUE GAME AVERAGES **HIGHER** ATTENDANCE RATES



COST TO ATTEND A GAME VARIES BY TEAM AND CITY

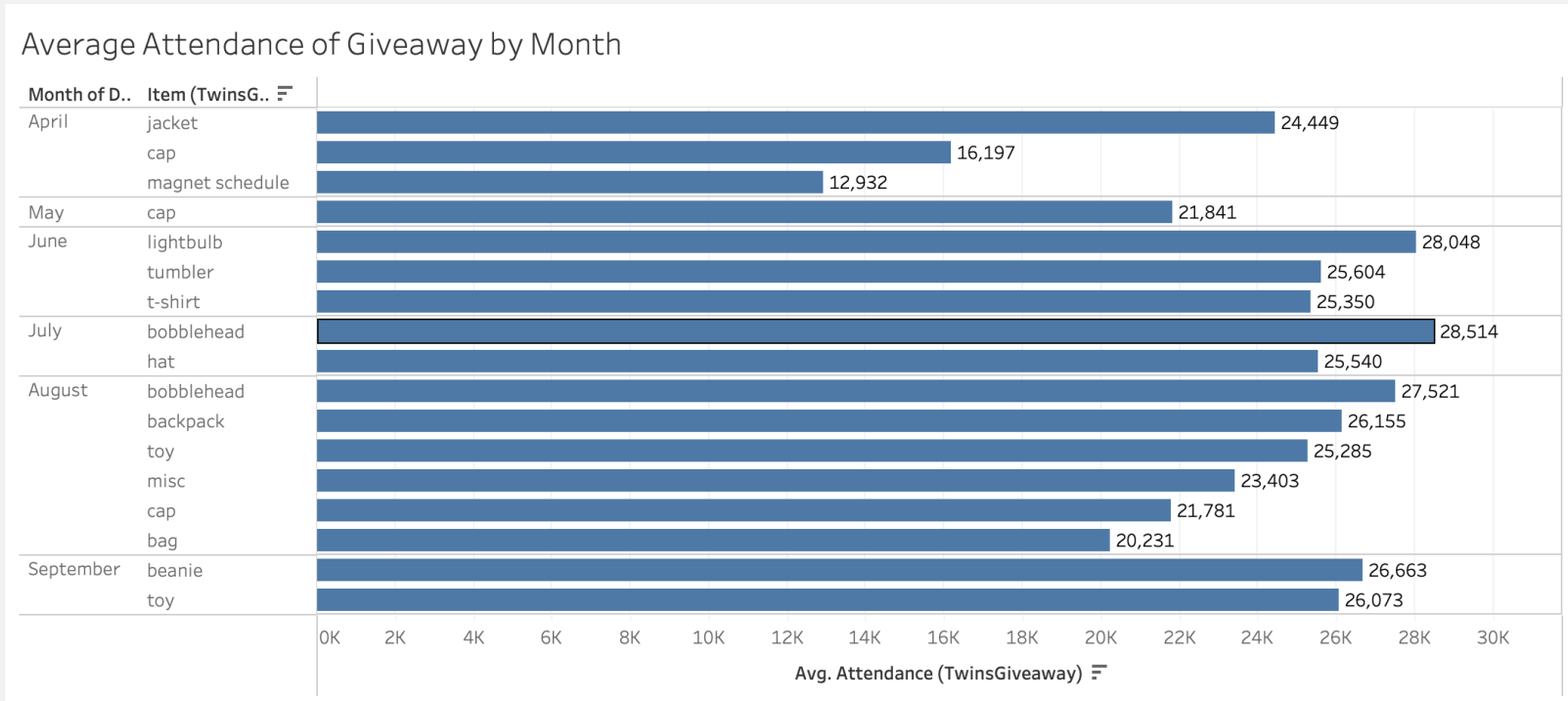
LEAGUE WIDE PROMOTIONAL TRENDS

- **STL Cardinals –** *Aggressive*
  - Widest giveaway opportunities
  - Clothing, bundled, Adult/16+ items average higher attendance
- **LA Dodgers –** *Passive*
  - Widest event opportunities
  - Bobbleheads highest count & sum of attendance
- **NY Yankees -** *Strategic*
  - Marketed 'collectible' items reflect game & event experiences
  - Differentiated limited edition items



TWINS PROMOTIONAL CLIMATE

Heavy saturation in summer months	Increase Spring/early Summer giveaway opportunities
Bobbleheads utilized ineffectively	Increase bobblehead frequency & variety
Higher attendance at child-specific giveaways	Explore & offer more family-oriented promotions



PROJECT RECOMMENDATIONS

April, May, September

PROJECT RECOMMENDATIONS

April, May, September

Spring Games

Spring Games

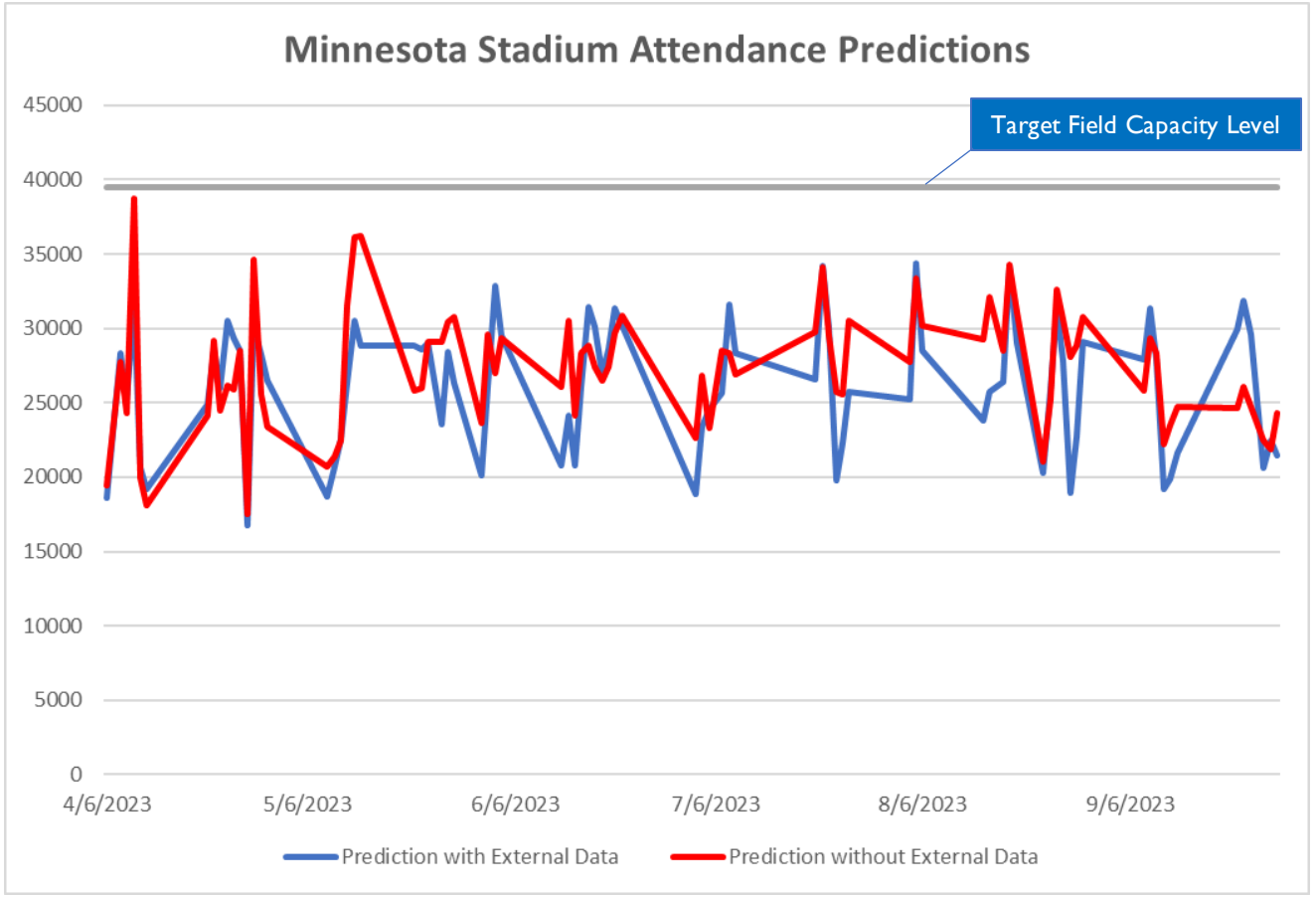
Spring Games

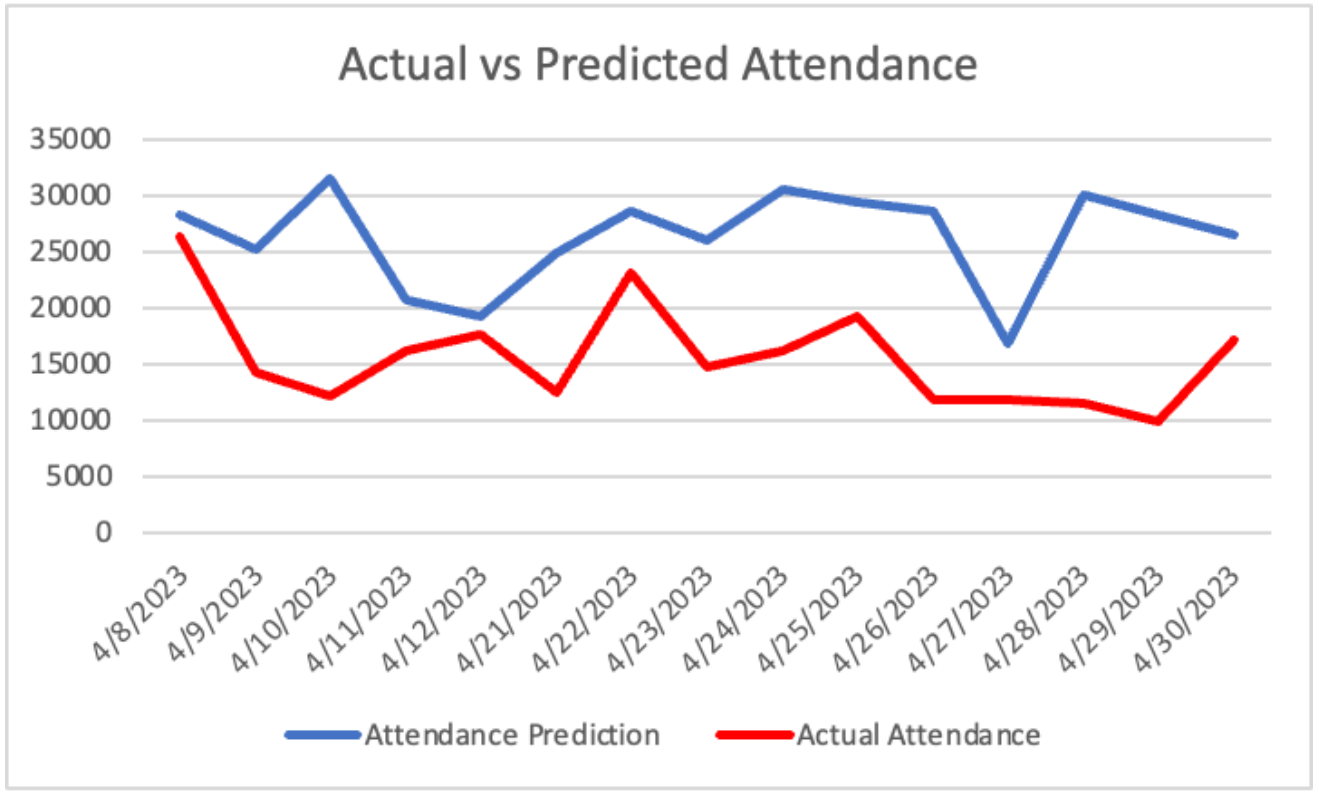
UTILIZING REGRESSION MODELS TO PREDICT 2023 MLB STADIUM ATTENDANCE

- Eight regression tests ran
- **Trained** on 20 years of data, **tested** on the 2022 season
- Evaluated on Mean Absolute Error and Mean Absolute Percentage Error
- **Random Forest** – All predictors without Cross Validation had the strongest results

Data Set	Model Approach	Train MAE	Test MAE	Difference	MAPE
Original Data	RF - All predictors wo CV	1951.817	2041.051	89.234	152.3%
External Data	RF - All predictors wo CV	1871.633	2239.578	367.945	80.3%
External data with subset data (previous 5 years)	RF - All predictors wo CV	1704.678	2034.616	329.938	62.9%
External Data with Previous Attendance (Dynamic Model)	RF - All predictors wo CV	1579.74	1384.599	-195.141	**21.2%**

STADIUM PREDICTION RESULTS

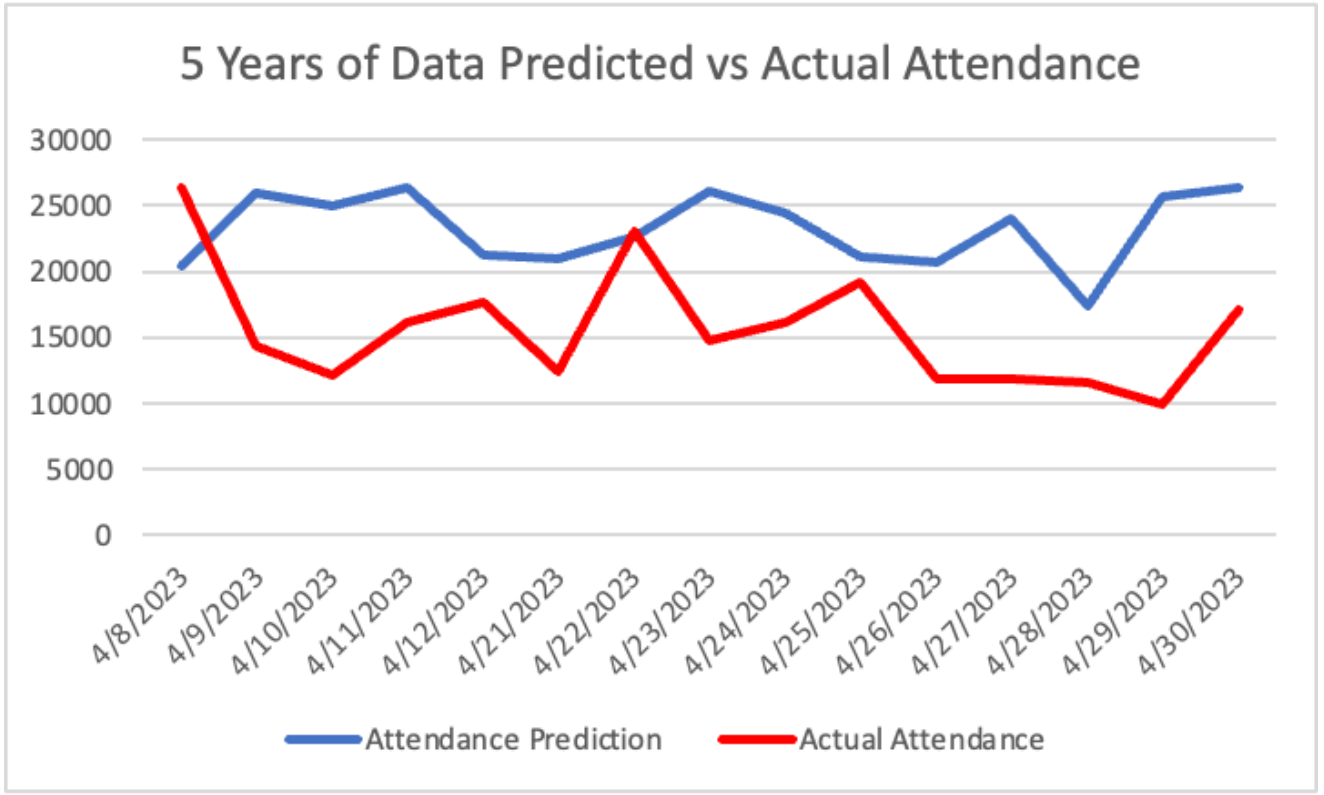




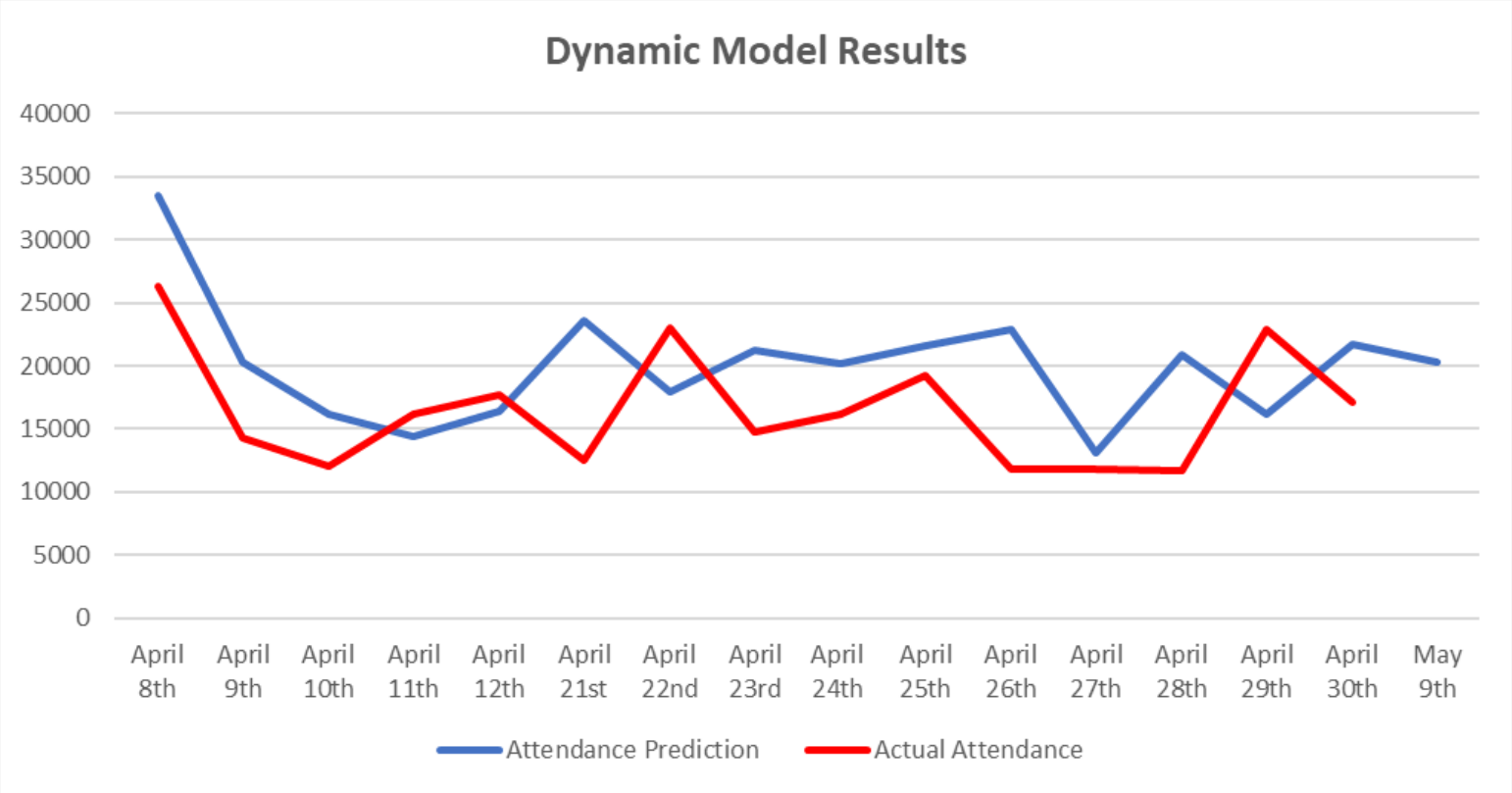
TWINS ATTENDANCE MODEL RESULTS FROM PLAYED GAMES

- External Data Model Predictions
- *4/9 was Easter – not noted in model

SMALLER SUBSET OF DATA LEADS TO BETTER PREDICTIONS



THE **DYNAMIC MODEL** PRODUCES THE MOST ACCURATE RESULTS



OUTCOMES AND STUDENT TAKEAWAYS

Communication is the key factor to team success

OUTCOMES AND STUDENT TAKEAWAYS

Communication is the key factor to team success

Being thorough when going through original data is crucial

OUTCOMES AND STUDENT TAKEAWAYS

Communication is the key factor to team success

Being thorough when going through original data is crucial

Setting a standard meeting time leads to higher quality of work

APPENDIX

1. St. Louis Cardinals Promotional Schedule, 2022
2. Los Angeles Dodgers Promotional Information, 2022
3. Los Angeles Dodgers Promotional, Cont.
4. New York Yankees Promotional Schedule, 2022
5. Minnesota Twins Promotional Information, 2022
6. Minnesota Twins Promotional Information, Cont.
7. Minnesota Twins Name History
8. Target Field Information

Minnesota Twins Attendance Growth Strategy

MLB 03-1: Sarah Philips, Katie Zawoyski, Beth Tiggelaar, Ryan Yuson

1.) Objective

The objective of this project was to build a predictive attendance model for MLB home games using historical game log and scheduling data, as well as provide actionable insight into underlying factors that influence home game attendance. Overarchingly, we aim to provide recommendations to positively impact attendance, as well as inform operational, revenue, and financial strategy impacts for the Minnesota Twins.

2.) Methodology

We first created a baseline regression model using our cleaned, initial data. From here, we explored Variable Importance through T-testing to helped determine what fields have a significant positive and negative impact on attendance.

Modeling – Regression

- Trained on 20 years of data, tested on the 2022 season
- Ran 8 different models and evaluated them on Mean Absolute Error
- Random Forest Model – All predictors without cross validation had the strongest results

3.) Data Preparation

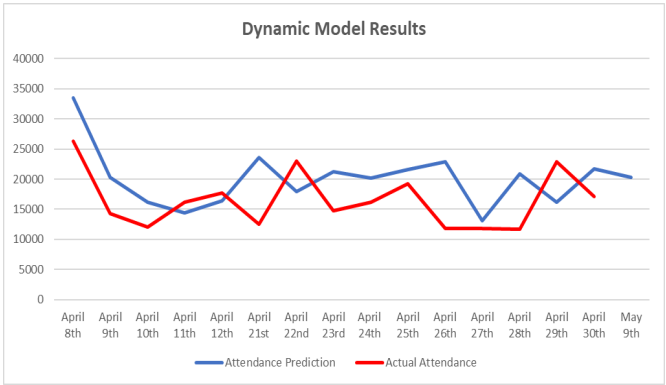
When working with the original datasets, our first step was to properly format all datatypes. While exploring the data, we created dummy variables for categoric features. The next step was to remove any instances of with null/negative attendance. Once the data was cleaned, we were able to implement external data and then finally subset the data for testing.

4.) Data Description

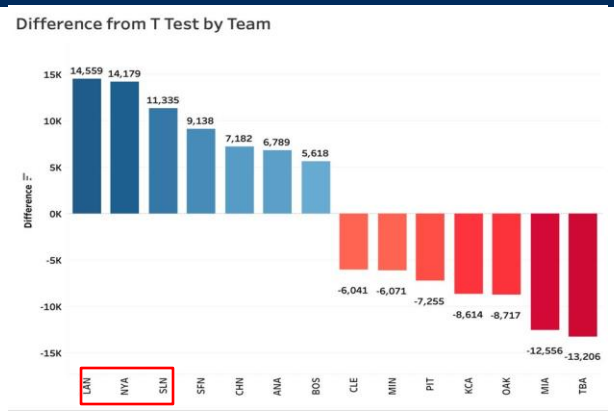
Our team was provided three csv files containing twenty years of historical game logs and scheduling data, as well as the current 2023 MLB schedule. Our first set of clean data consisted of 263 variables across 54,345 instances. The further use of external data provided 6 additional variables for modeling: stadium capacity, city population, number of professional teams/city, cost to attend a game, and previous game attendance of the current 2023 season. Promotional scheduling from the 2022 season provided information regarding free/exclusive giveaway items, supply limits, game themes, and pre/post game events. It is important to note that this data was used strictly for exploration as it is not fully comprehensive of league scheduling.

5.) Regression Results and Findings

To predict 2023 MLB attendance, we used R studio to run a random forest all predictors without cross validation to run a regression to get the 2023 predictions. We ran this test three times with: 1.) original data, 2.) integrated external data, and 3.) integrated previous game attendance. We found that the dynamic model using each game's previous attendance made the most accurate predictions, exhibiting the lowest test MAE (1,384.59) and MAPE (21.2%), while remaining considerably tolerant when exposed to new data.



The following t-test displays the teams with the most positive and negative impact on home game attendance. The positive impact of the LA Dodgers, NY Yankees, and STL Cardinals presents the opportunity to explore event/giveaway schedules across the league and provide insight into trends and successful promotional strategies.



6.) Business Recommendations

After conducting variable exploration, we found the variables that had low attendance indicate for the twins to target the months of April and September, weekdays including Tuesdays – Thursdays, and more general spring and league games. To increase attendance in these areas we recommend utilizing a variety of complemented promotional activities including themed games, giveaways, ticket bundles, and weekday promotions to add perceived value to games without the insistence of significant additional costs attend.

Promotional Placement

Exploring 2022 promotional scheduling gave insight into a wide range of successful promotional strategies. The Minnesota Twins should restructure their approach by more opportunities in first half of the season, including the expansion of offered items and incorporation of additional themes and events alongside items to mitigate stand-alone 'exclusive' giveaways.