Julia Sveen

Case Study 1: “Analysis of Percent Capacity at Dodgers Stadium from 2014-2022”

February 27th, 2024

Introduction:

Baseball has been a major part of my life since I can possibly remember. Growing up, I played a variety of sports, but baseball gained the bulk of my fascination. My father definitely played a role in this and has taken me to my fair share of major league baseball games across the country. Out of all teams, the Dodgers and Dodgers Stadium remains my favorite, (although it might be tied with Fenway Park). For this analysis, I want to determine the factors that have an impact on attendance at Dodgers Stadium in the scope of percent capacity; such as opponent, month, team performance, and more. I hope to gain insight into different elements that contribute to an understanding of fan behavior with the end goal of optimizing fan experience at the famous Dodgers Stadium, as well as contributing to the team’s profit.

Method:

Data Gathering:

We were tasked with reading in two main datasets for the analysis: MLB\_Attendance, which contains attendance data for MLB games, and MLB\_Venue\_Capacity, which provides information on the capacity of each stadium.

Variables Examined:

Percent Capacity (Dependent Variable): I utilized data from the venue capacity dataset and merged it to the attendance dataset to calculate the percentage of venue capacity (Pct\_Cap) as the outcome variable (y-variable).

Independent Variables - I examined various factors that could influence attendance, including:

* Team performance indicators (e.g., games back, win percentage)
* Game-specific variables (e.g., day of the week, month of the year)
* Certain opponents (rivalries, teams played that typically indicate successful regular seasons)
	+ I determined which opponents to test in my regression models based on the opponents that generated the highest percent capacity on average (see Fig 2). I used the Houston Astros, Los Angeles Angels, and the Boston Red Sox, as well as the San Francisco Giants, and most of those teams actually ended up being statistically significant in my regression analysis.

Data Cleaning and Preparation: I cleaned and preprocessed the data, including converting categorical variables into factors, joining datasets on relevant variables, adding helpful rows, filtering by team, and more.

Model Building: I built three different regression models to predict attendance using different combinations of independent variables. I then evaluated the models based on their R-squared values, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Results:

Model 1:

Outcome Variable: Pct\_Cap

Independent Variables: GB\_Pre, WPct\_Pre, DayofWeek\_Friday, DayofWeek\_Saturday, Month\_June, Month\_July, Month\_August, Month\_September, Month\_October

R-squared: 0.1655

MAE: 0.09018638

RMSE: 0.01457584

Model 2:

Outcome Variable: Pct\_Cap

Independent Variables: RS\_PG, RA\_PG, Opp\_RS\_PG, Opp\_RA\_PG, Opp\_SFG, Opp\_HOU, Opp\_LAA

R-squared: 0.1016

MAE: 0.09318998

RMSE: 0.01394638

Model 3:

Outcome Variable: Pct\_Cap

Independent Variables: RS\_PG, RA\_PG, Opp\_RA\_PG, OppWPct\_Pre, DayofWeek\_Saturday, Month\_June, Month\_July, Month\_August, Month\_September, Month\_October, Opp\_SFG, Opp\_LAA

R-squared: 0.2186

MAE: 0.09414765

RMSE: 0.00887246

Discussion of Findings:

All three of my models show fairly moderate to good performance in predicting attendance, as indicated by their R-squared values. However, I think that Model 3 overall is the most accurate in predicting attendance (in the scope of percentage capacity) because it has the highest R-squared value out of all three of my models. An R-squared value of 0.2186 indicates that the explanatory variables explained about 22% of the variation in percent capacity. Although the MAE value is 0.09414765 (and is slightly higher than my other models’), the value is not significantly different and suggests that my model is off by about 9% in terms of percent capacity. For my first model, I focused mostly on the factored data (the month and day of the week). For my second, I focused mainly on more statistical-related factors, particularly ones to do with the opponent. For the third, I combined many of the statistically significant factors from the first two models. In general, statistics pertaining to the away team (the opponent), the months from July-October, and the number of runs scored/runs allowed going into the game are all important factors in determining the percent capacity at Dodgers Stadium. These findings can help the Dodgers organization in creating effective marketing strategies, optimizing their schedule, as well optimizing ticket sales to maximize attendance at games.

Appendix:

Fig 1

Fig 2

I was able to derive this bar chart by using ‘dummy coding’ based on the Dodgers’ opponent (OPP variable). After doing this, I arranged the dataset in descending order and graphed just the teams in the top third of percent capacity for better readability.

R Code:

*# case study 1*

*# Julia Sveen*

*# library declarations*

library(readr)

library(dplyr)

library(ggplot2)

library(readxl)

library(fastDummies)

library(caTools)

*# gathering of files*

MLB\_Attendance=read.csv("MLB\_Attendance.csv")

View(MLB\_Attendance)

MLB\_Venue\_Capacity <- read\_excel("MLB\_Venue\_Capacity.xlsx")

View(MLB\_Venue\_Capacity)

*# cleaning/mutating data*

MLB\_Attendance=MLB\_Attendance%>%mutate(DayofWeek=as.factor(DayofWeek))

MLB\_Attendance=MLB\_Attendance%>%mutate(Month=as.factor(Month))

Final\_Att=left\_join(MLB\_Attendance,MLB\_Venue\_Capacity,by=c("venue","Year")) %>%

  mutate(Pct\_Cap=Att/cap)

*# converting char vars into factored vars for leveling*

Final\_Att=Final\_Att%>%mutate(DayofWeek=factor(DayofWeek,levels=c("Monday",

                                                                 "Tuesday",

                                                                 "Wednesday",

                                                                 "Thursday",

                                                                 "Friday",

                                                                 "Saturday",

                                                                 "Sunday")),

                             Month=factor(Month,levels=c("March",

                                                         "April",

                                                         "May",

                                                         "June",

                                                         "July",

                                                         "August",

                                                         "September",

                                                         "October")))

Att\_Dummied=dummy\_cols(Final\_Att,select\_columns=c("DayofWeek","Month","Opp"),

                       remove\_first\_dummy = TRUE)

*# filtering to just LA Dodgers*

Dodgers\_Att=Att\_Dummied %>% filter(Tm=="LAD")

*# bar chart describing relationship b/w day of the week and % capacity*

ggplot(Dodgers\_Att,aes(x=DayofWeek,y=Pct\_Cap))+geom\_col(fill="royalblue")+

         labs(x="Day of the Week",y="% Capacity")+

  ggtitle("Relationship between Day of the Week and Percent Capacity at

          Dodgers Stadium")

*# barchart of % capacity based on certain opponents*

Avg\_capacity\_per\_team<-Dodgers\_Att %>% group\_by(Opp) %>%

  summarise(Avg\_pct\_cap=mean(Pct\_Cap))

Avg\_capacity\_per\_team <- Avg\_capacity\_per\_team %>%arrange(desc(Avg\_pct\_cap))

total\_rows<-nrow(Avg\_capacity\_per\_team)

top\_third\_rows<-ceiling(total\_rows/3)

top\_third\_data <- Avg\_capacity\_per\_team[1:top\_third\_rows, ]

ggplot(top\_third\_data,aes(x=Opp,y=Avg\_pct\_cap))+

  geom\_col()+labs(x="Opponent",y="Average % Capacity")+

  ggtitle("Top Third Average % Capacity According to Opponent")

*# I'm highlighting 3 teams because 2 are indicative of being in the post season*

  *# Red Sox & Astros*

*# one is indicative of very close proximity (Angels)*

set.seed(123)

Sample=sample.split(Dodgers\_Att$Pct\_Cap,SplitRatio=.75)

LAD\_Train=subset(Dodgers\_Att,Sample==TRUE)

LAD\_Test=subset(Dodgers\_Att,Sample==FALSE)

*# build basic model*

*# model 1*

LAD\_Model\_1=lm(Pct\_Cap~GB\_Pre+WPct\_Pre+

                 DayofWeek\_Friday+DayofWeek\_Saturday+

                Month\_June+Month\_July+Month\_August+

                 Month\_September+Month\_October,data=LAD\_Train)

summary(LAD\_Model\_1)

*# Adjusted R^2 = 0.1475*

LAD\_Test$Pred\_1=predict(LAD\_Model\_1,newdata = LAD\_Test,type="response")

mean(abs(LAD\_Test$Pct\_Cap-LAD\_Test$Pred\_1))

*# 0.09018638*

sqrt(mean(LAD\_Test$Pct\_Cap-LAD\_Test$Pred\_1)^2)

*# 0.01457584*

*# model 2*

LAD\_Model\_2=lm(Pct\_Cap~RS\_PG+RA\_PG+Opp\_RS\_PG+Opp\_RA\_PG+Opp\_SFG+

                 Opp\_HOU+Opp\_LAA,data=LAD\_Train)

summary(LAD\_Model\_2)

*# Adjusted R^2 = 0.08651*

LAD\_Test$Pred\_2=predict(LAD\_Model\_2,newdata = LAD\_Test,type="response")

mean(abs(LAD\_Test$Pct\_Cap-LAD\_Test$Pred\_2))

*# 0.09318998*

sqrt(mean(LAD\_Test$Pct\_Cap-LAD\_Test$Pred\_2)^2)

*# 0.01394638*

*# model 3*

LAD\_Model\_3=lm(Pct\_Cap~RS\_PG+RA\_PG+Opp\_RA\_PG+OppWPct\_Pre+DayofWeek\_Saturday+

                 Month\_June+Month\_July+Month\_August+

                 Month\_September+Month\_October+Opp\_SFG+Opp\_LAA,data=LAD\_Train)

summary(LAD\_Model\_3)

*# Adjusted R^2 = 0.1959*

LAD\_Test$Pred\_3=predict(LAD\_Model\_3,newdata = LAD\_Test,type="response")

mean(abs(LAD\_Test$Pct\_Cap-LAD\_Test$Pred\_3))

*# 0.09414765*

sqrt(mean(LAD\_Test$Pct\_Cap-LAD\_Test$Pred\_3)^2)

*# 0.00887246*