3.2 Assignment

Scott Breitbach

9/13/2021

3.2 Assignment: Using Data to Improve a Marketing Promotion

For this week's assignment we're going to use Dodgers Major League Baseball data from 2012. The data file you will be using is contained in the dodgers.csv file. I would like you to determine what night would be the best to run a marketing promotion to increase attendance. It is up to you if you decide to recommend a specific date or if you recommend a day of the week (e.g., Tuesdays) or month and day of the week (e.g., July Tuesdays). Use R and/or Python to accomplish this assignment. It is important to remember, there will be lots of ways to solve this problem. Explain your thought process and how you used various techniques to come up with your recommendation.

```
# Load CSV file to Data Frame
df <- read.csv("dodgers.csv", header=TRUE)

# Get a preliminary look at the data
head(df)
```

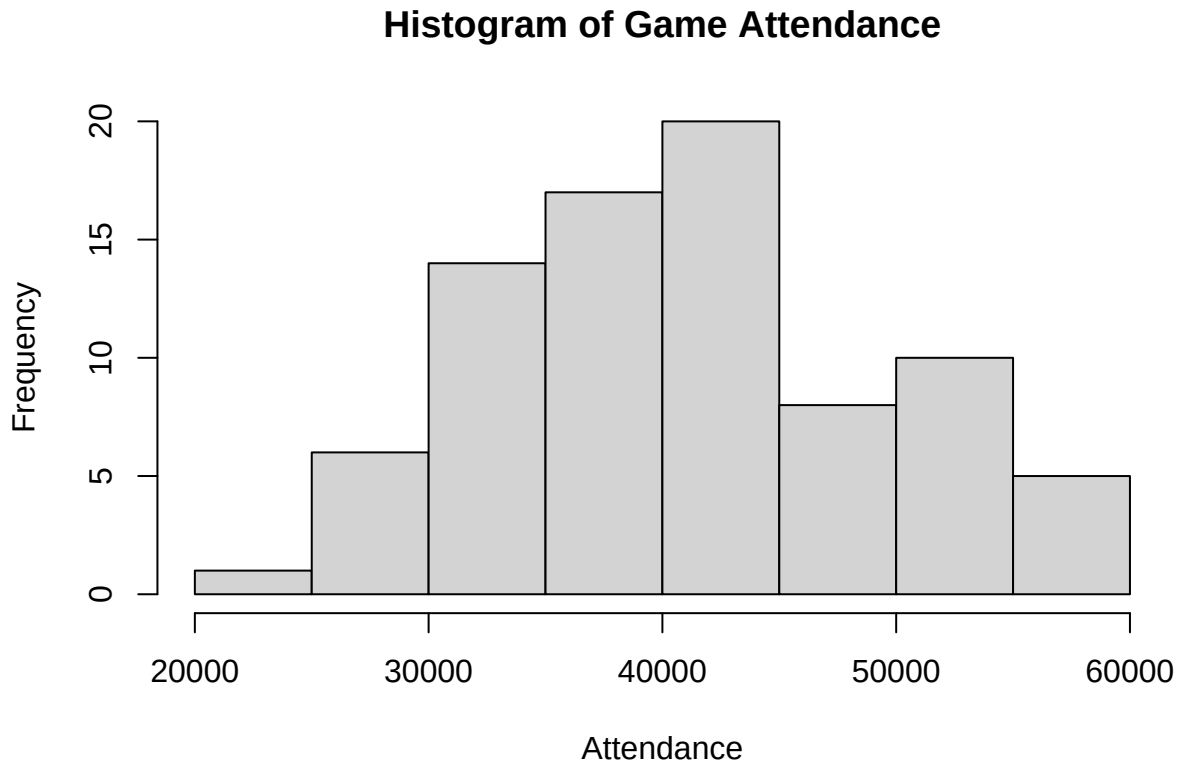
```
##   month day attend day_of_week opponent temp  skies day_night cap shirt
## 1   APR  10  56000     Tuesday  Pirates   67 Clear        Day  NO    NO
## 2   APR  11  29729   Wednesday  Pirates   58 Cloudy     Night  NO    NO
## 3   APR  12  28328    Thursday  Pirates   57 Cloudy     Night  NO    NO
## 4   APR  13  31601      Friday   Padres   54 Cloudy     Night  NO    NO
## 5   APR  14  46549    Saturday   Padres   57 Cloudy     Night  NO    NO
## 6   APR  15  38359      Sunday   Padres   65 Clear        Day  NO    NO
##   fireworks bobblehead
## 1        NO         NO
## 2        NO         NO
## 3        NO         NO
## 4       YES         NO
## 5        NO         NO
## 6        NO         NO
```

```
summary(df)
```

```
##     month                day             attend      day_of_week
##  Length:81          Min.   : 1.00   Min.   :24312   Length:81
##  Class :character   1st Qu.: 8.00   1st Qu.:34493   Class :character
##  Mode  :character   Median :15.00   Median :40284   Mode  :character
##                     Mean   :16.14   Mean   :41040
##                     3rd Qu.:25.00   3rd Qu.:46588
##                     Max.   :31.00   Max.   :56000
```

```
##    opponent              temp          skies            day_night
##  Length:81          Min.   :54.00   Length:81          Length:81
##  Class :character   1st Qu.:67.00   Class :character   Class :character
##  Mode  :character   Median :73.00   Mode  :character   Mode  :character
##                     Mean   :73.15
##                     3rd Qu.:79.00
##                     Max.   :95.00
##      cap                shirt            fireworks          bobblehead
##  Length:81          Length:81          Length:81          Length:81
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
# Create Histogram of Attendance since that is our key variable
hist(df$attend, xlab="Attendance", main="Histogram of Game Attendance")
```

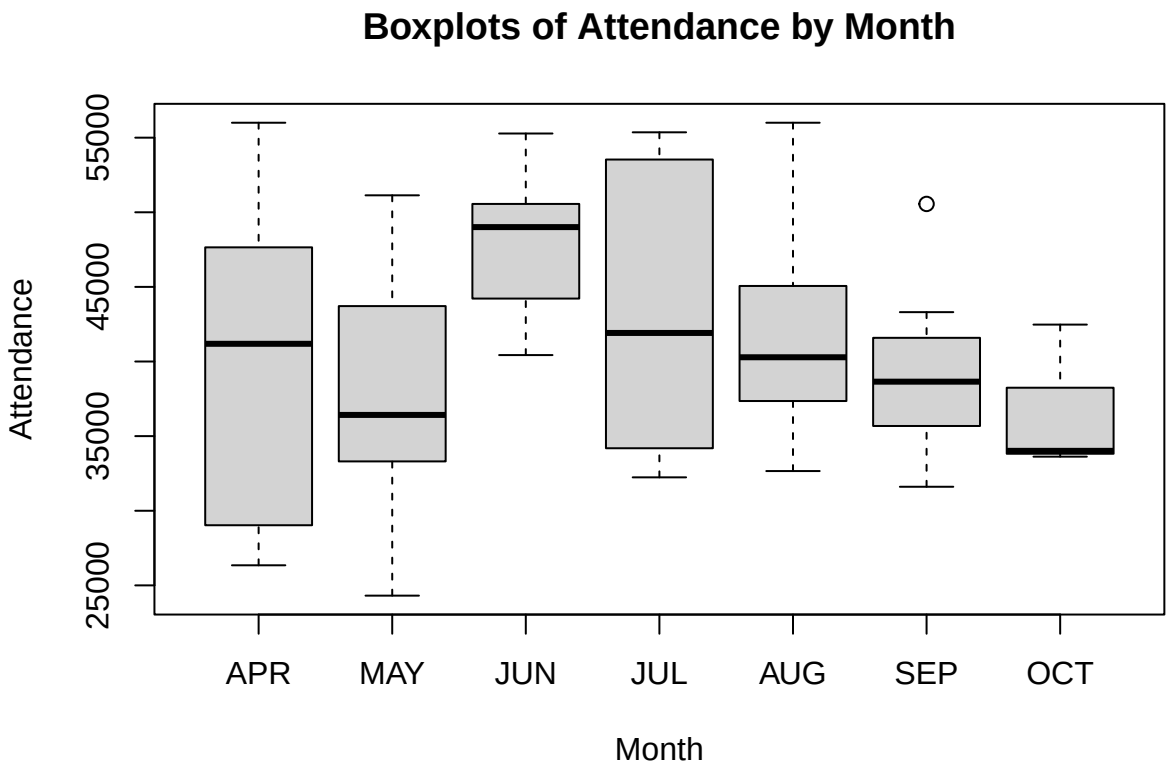


The mean/median attendance values are pretty close, around 40-41 thousand, which will be our rough benchmark. I'll go through the variables in order to look for variations and to see if there is any indication of correlation with attendance that can be exploited to guide marketing promotions.

First up: Month

```
# Assign month order so they chart correctly
df$month <- factor(df$month, levels=c("APR", "MAY", "JUN", "JUL",
```

```
                                        "AUG", "SEP", "OCT"))
# Create boxplot
boxplot(attend ~ month, df, xlab="Month", ylab="Attendance",
        main="Boxplots of Attendance by Month")
```



It appears the lowest months are May and October. If I had to choose one of those months to do a promotion, I would probably select May because it shows more variation, so a promotion may have a greater effect.

Up next: Day of the month. I wouldn't expect this to have any impact on attendance. If anything, one might see a slight uptick after the first and fifteenth of each month just because some people get paid on those days.

```
# First check for a correlation
print("Day/Attendance Correlation:")
```

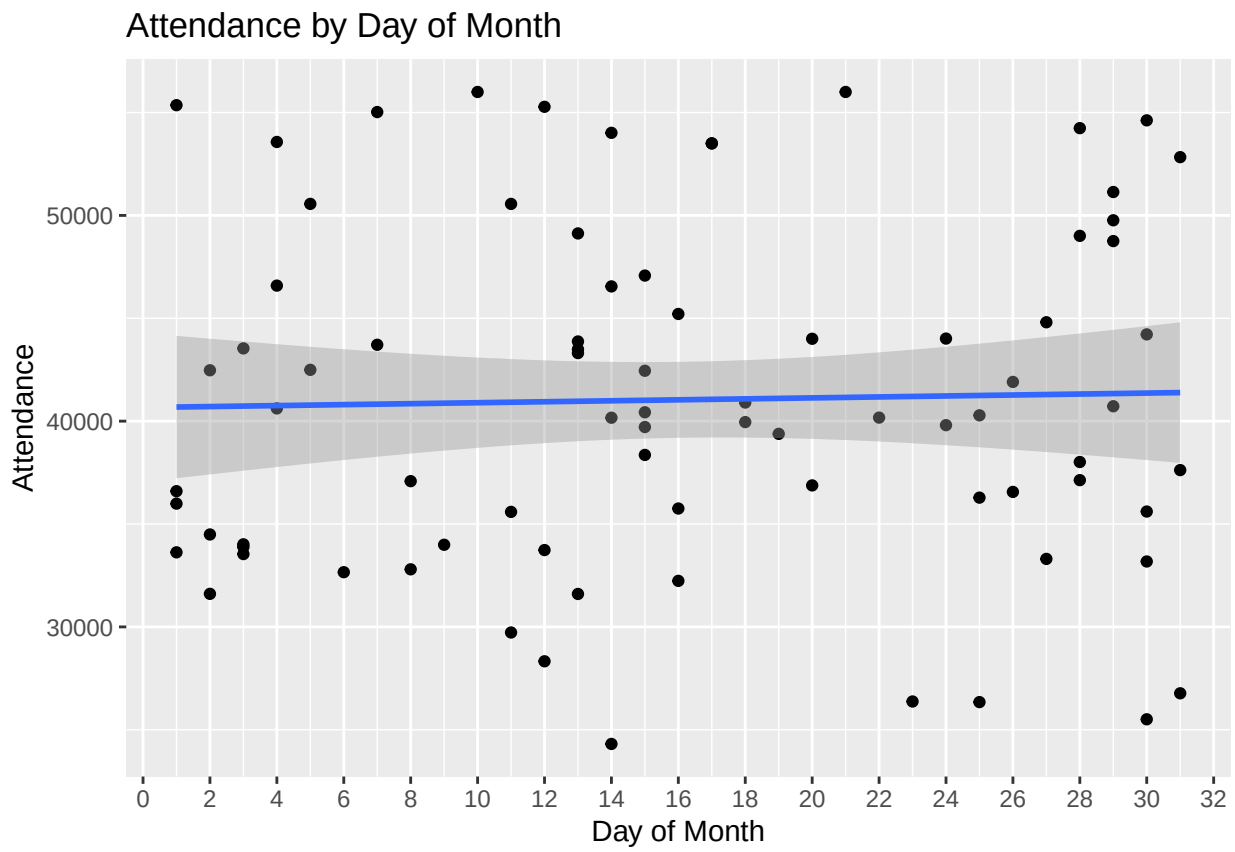
```
## [1] "Day/Attendance Correlation:"
```

```
cor(df$attend, df$day)
```

```
## [1] 0.02709298
```

```
# Plot attendance by day of month
ggplot(df, aes(x=day, y=attend)) + geom_point() + geom_smooth(method="lm") +
  labs(x="Day of Month", y="Attendance", title="Attendance by Day of Month") +
  scale_x_continuous(n.breaks=20)
```

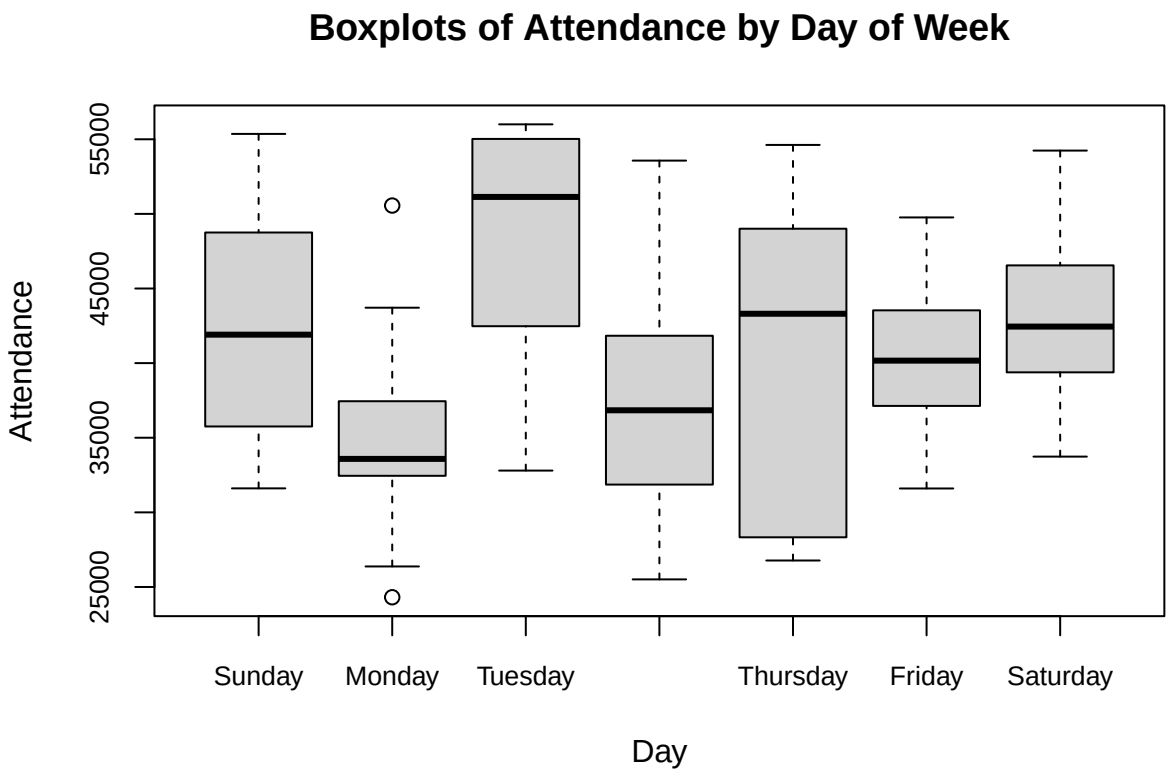
```
## `geom_smooth()` using formula 'y ~ x'
```



As expected, correlation is very low and there are no apparent patterns to the day of the month.

On to Day of the Week, where I expect to see some trends.

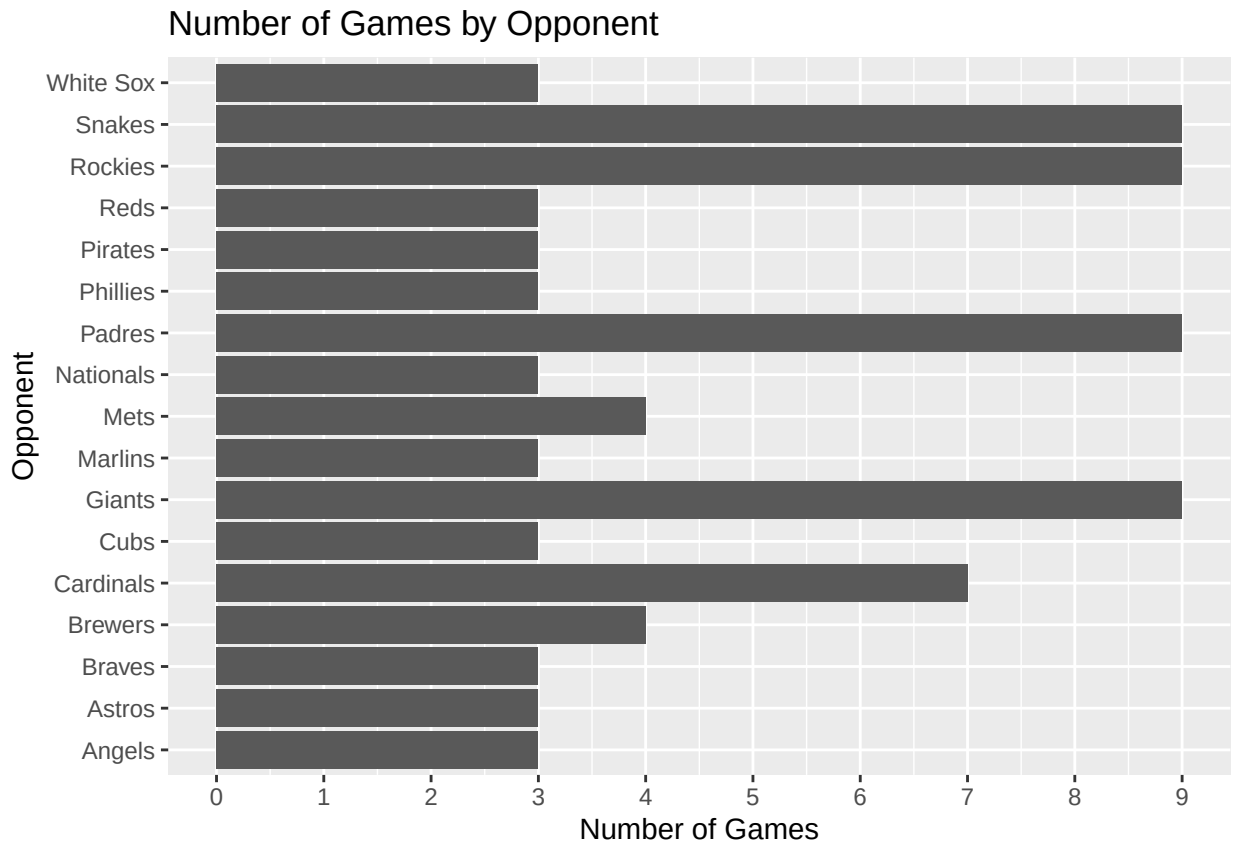
```
# Assign day order so they chart correctly
df$day_of_week <- factor(df$day_of_week,
                         levels=c("Sunday", "Monday", "Tuesday", "Wednesday",
                                  "Thursday", "Friday", "Saturday"))
# Create boxplot
boxplot(attend ~ day_of_week, df, xlab="Day", ylab="Attendance",
        main="Boxplots of Attendance by Day of Week", cex.axis=0.8)
```



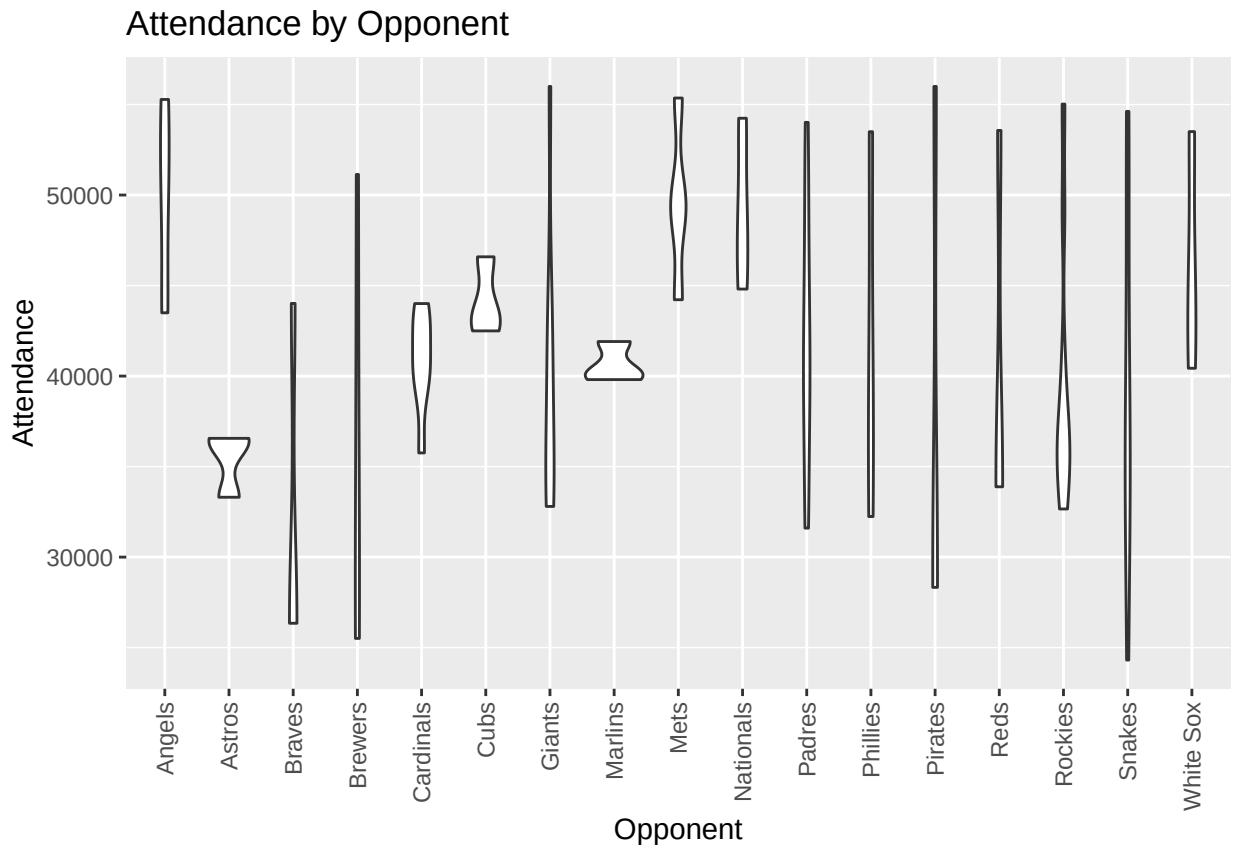
Mondays and Wednesdays appear to have room for improvement and may be good candidates for a promotion to increase attendance.

No to look at opponents. There are 17 teams and 81 games, which is fewer than 5 games per team on average, so I don't expect to see much useful information. Perhaps some teams are more popular or are big rivals and we might already expect to see higher attendance on those nights.

```
# Create bar plot showing number of games by opponent
ggplot(df, aes(x=as.factor(opponent))) + geom_bar() + coord_flip() +
  scale_y_continuous(n.breaks=8) + labs(x="Opponent", y="Number of Games",
                                        title="Number of Games by Opponent")
```




```
# Create violin plots of game attendance by opponent
ggplot(df, aes(x=opponent, y=attend)) + geom_violin() +
  labs(x="Opponent", y="Attendance", title="Attendance by Opponent") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

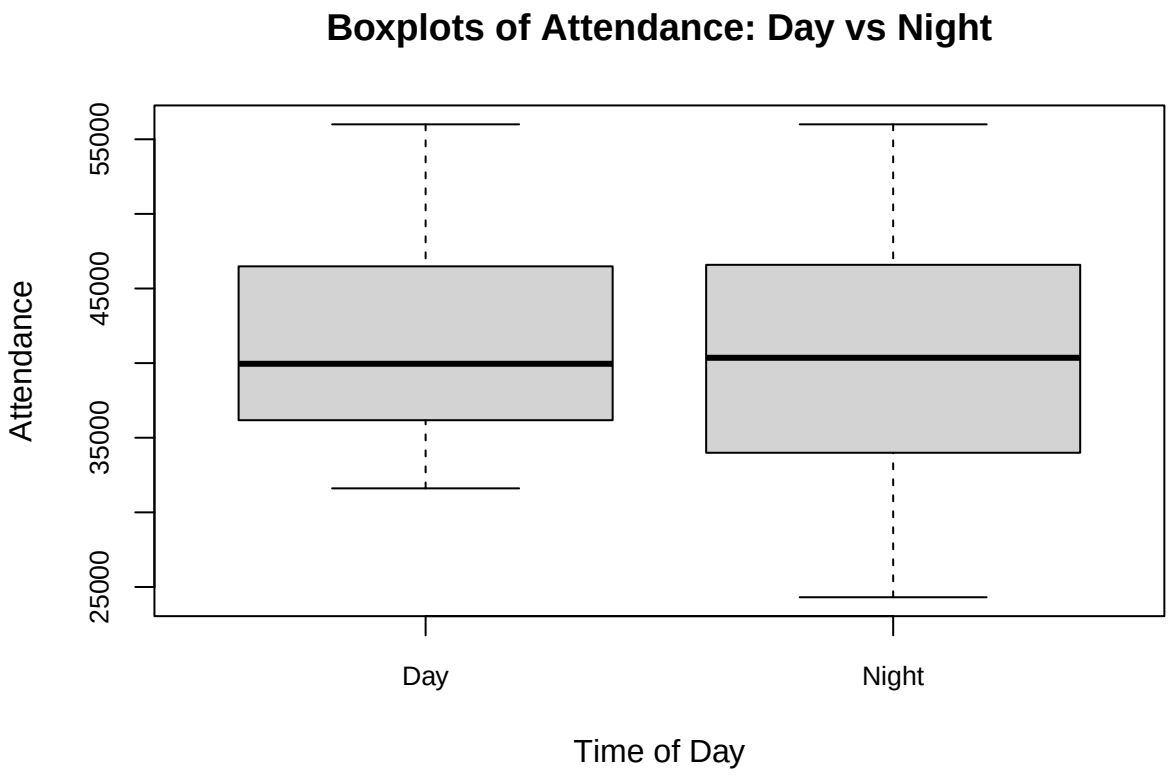


There are a handful of opponents that seem to have fairly consistent attendance. Of these, most have only 3-4 games so it's hard to draw conclusions. The only exception might be the Cardinals, who have a fairly consistent attendance over 7 games. Overall, I don't think the opponent should figure into a marketing promotion.

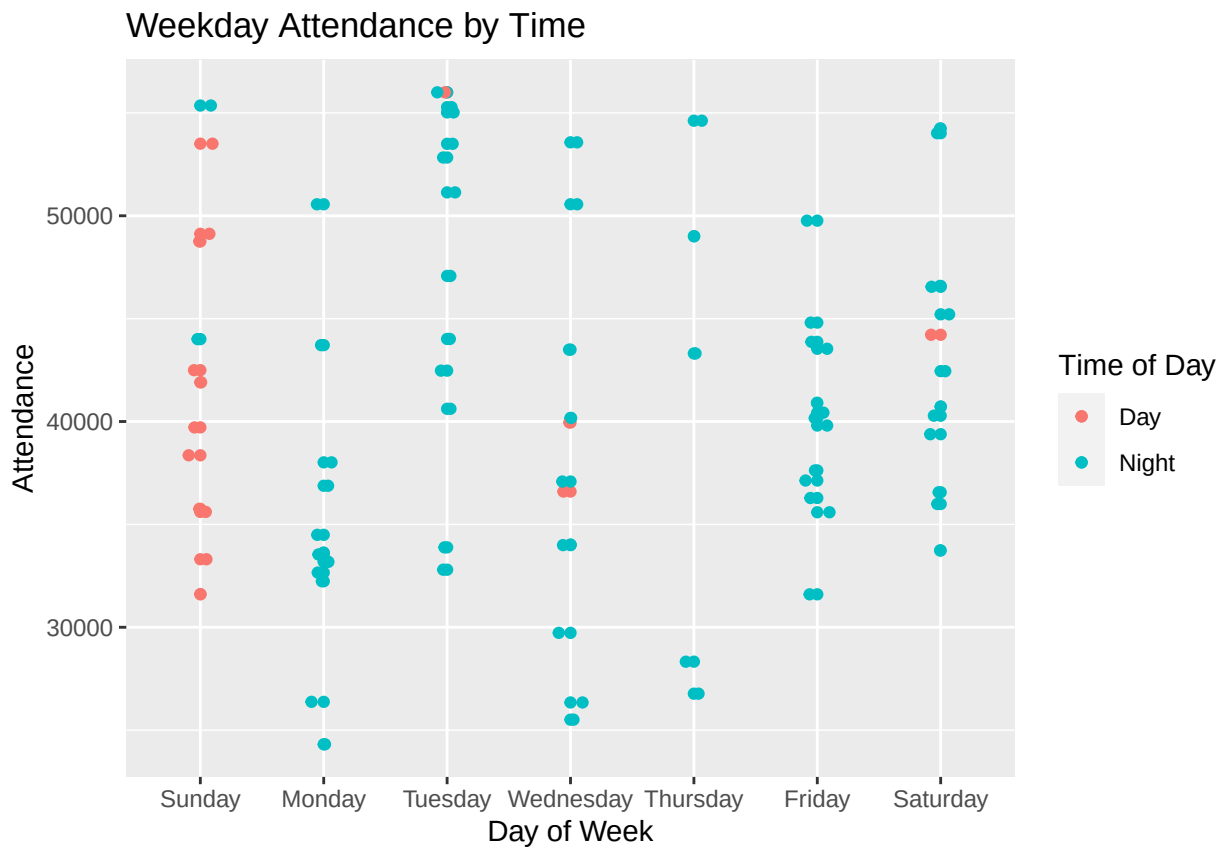
Because our goal is determining which night to hold a promotion, I'm going to skip over weather-related variables 'temp' and 'skies' because we cannot control or predict what the weather will be, apart from seasonal changes, which should be reflected in the 'month' variable.

I don't know what we'll see with day games vs night games.

```
# Boxplot of attendance for day vs night games
boxplot(attend ~ day_night, df, xlab="Time of Day", ylab="Attendance",
        main="Boxplots of Attendance: Day vs Night", cex.axis=0.8)
```



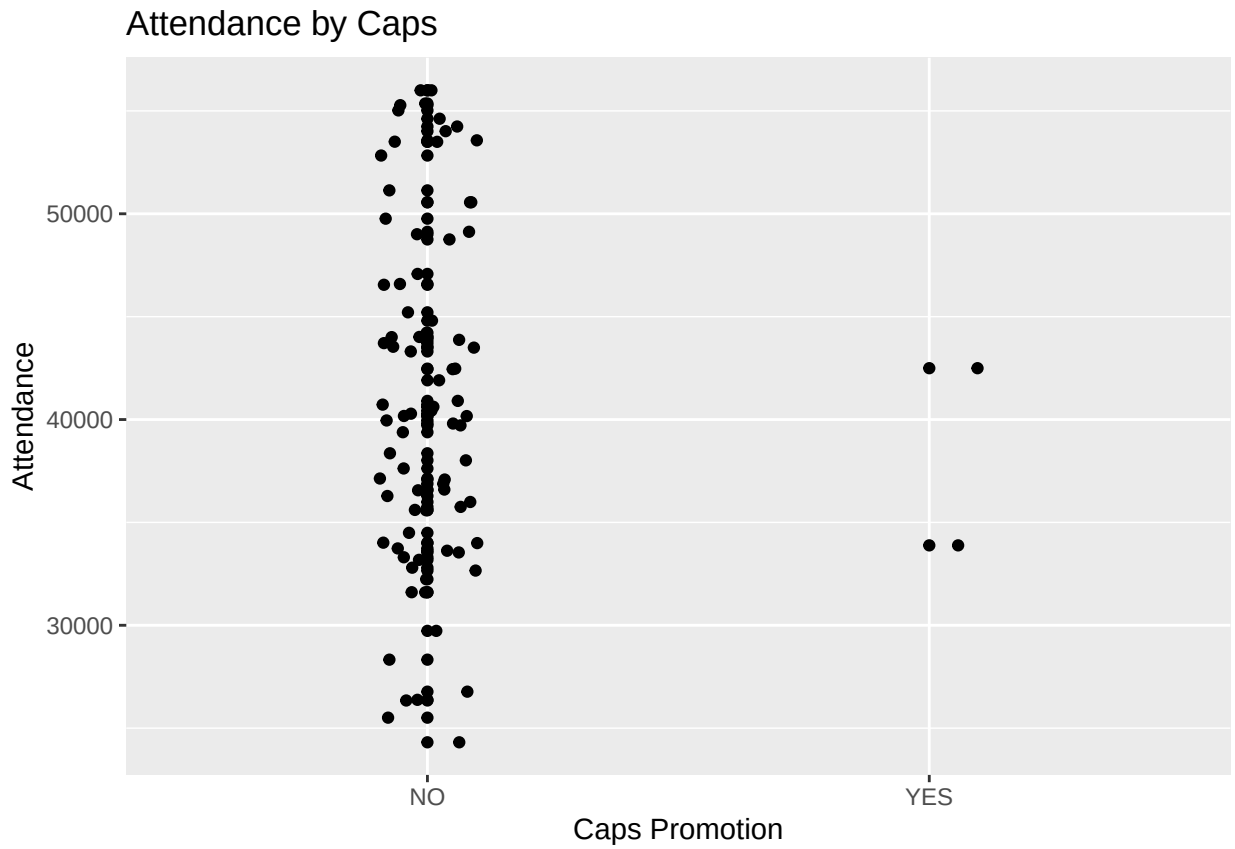

```
# Plot day vs night games by day of week
ggplot(df, aes(x=day_of_week, y=attend, color=day_night)) + geom_point() +
  labs(x="Day of Week", y="Attendance", title="Weekday Attendance by Time") +
  geom_jitter(width=0.1) + guides(color=guide_legend(title="Time of Day"))
```



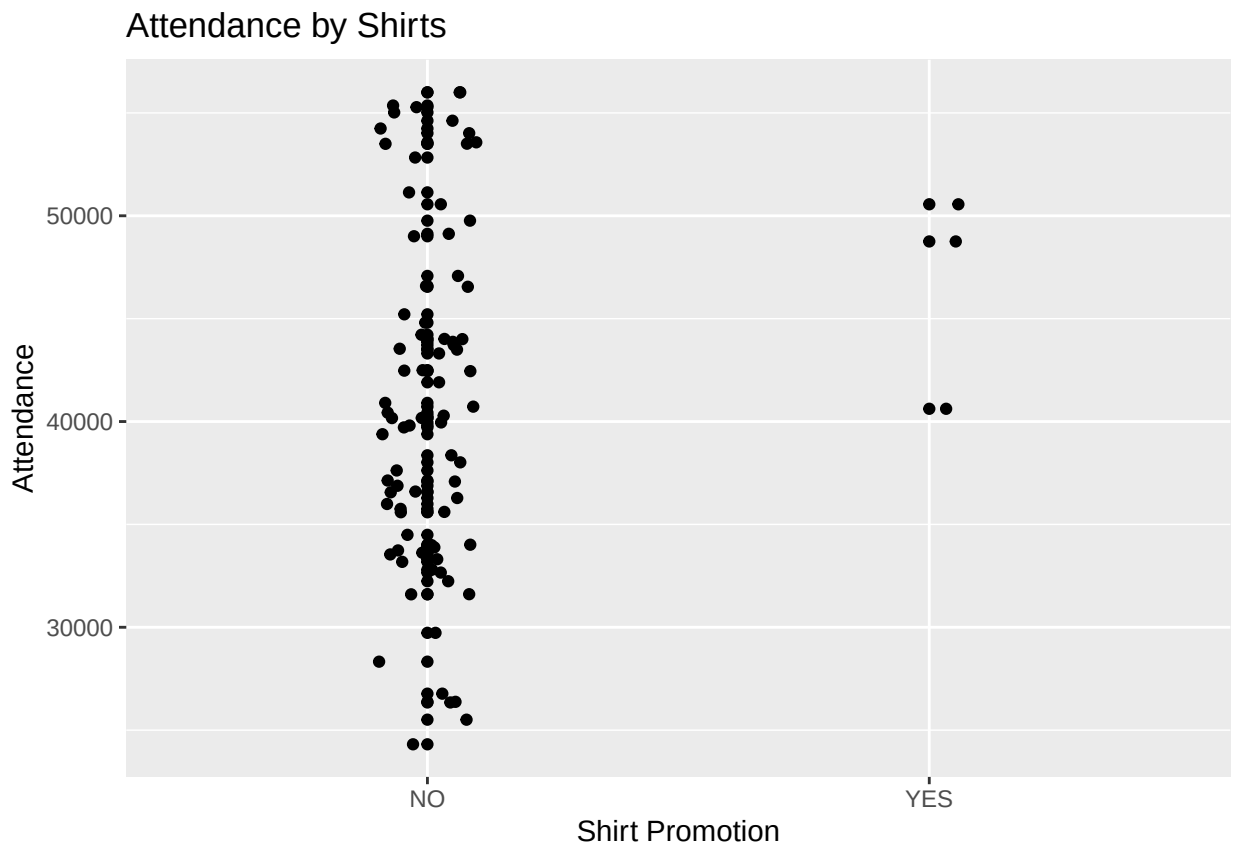
There doesn't appear to be much difference in attendance between day and night games and it appears that a decent rule of thumb would be that Day games happen on Sundays, while the others are Night games. Given that we're mostly going to be looking at Mondays and Wednesdays anyway, the time of the game should not be taken into account for the promotion.

Other promotions: caps, shirts, fireworks, and bobbleheads.

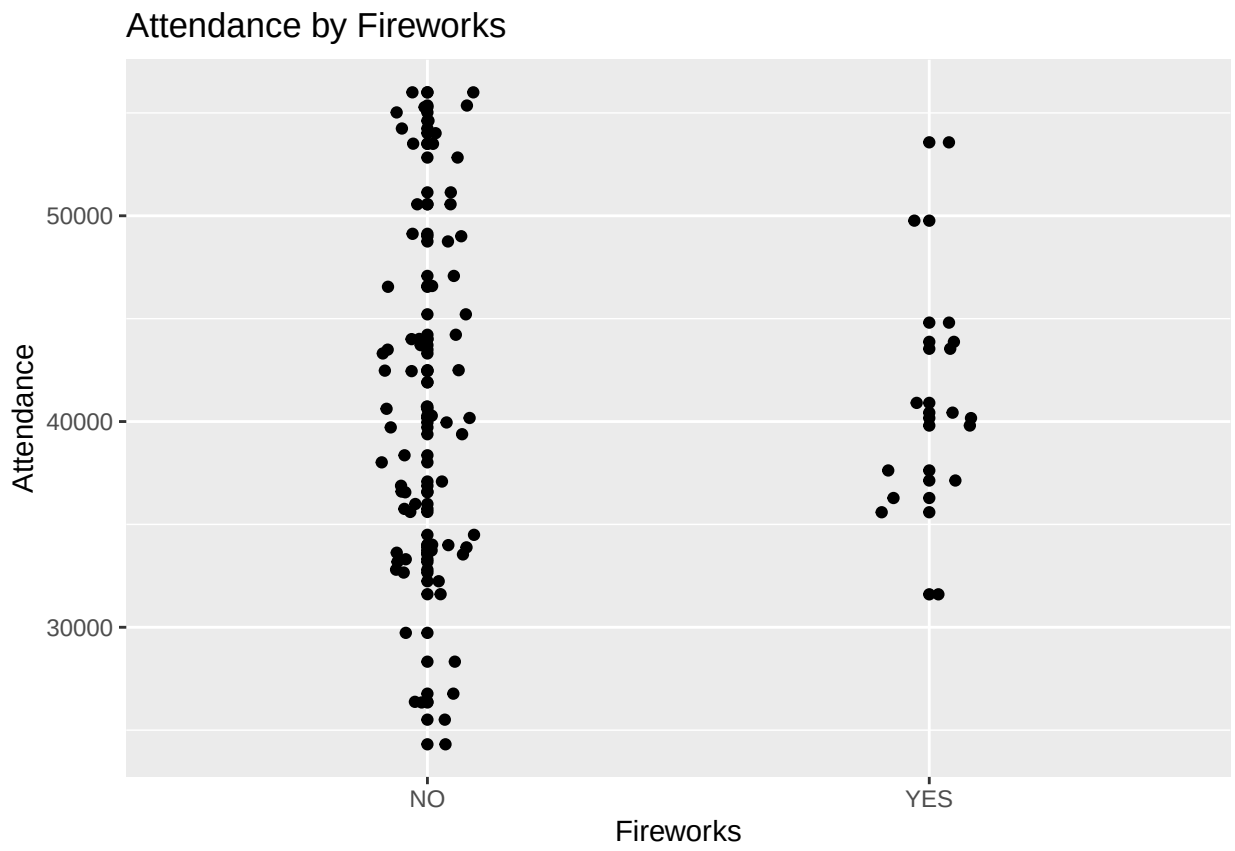
```
# Plot caps vs attendance
ggplot(df, aes(x=cap, y=attend)) + geom_point() + geom_jitter(width=0.1) +
  labs(x="Caps Promotion", y="Attendance", title="Attendance by Caps")
```



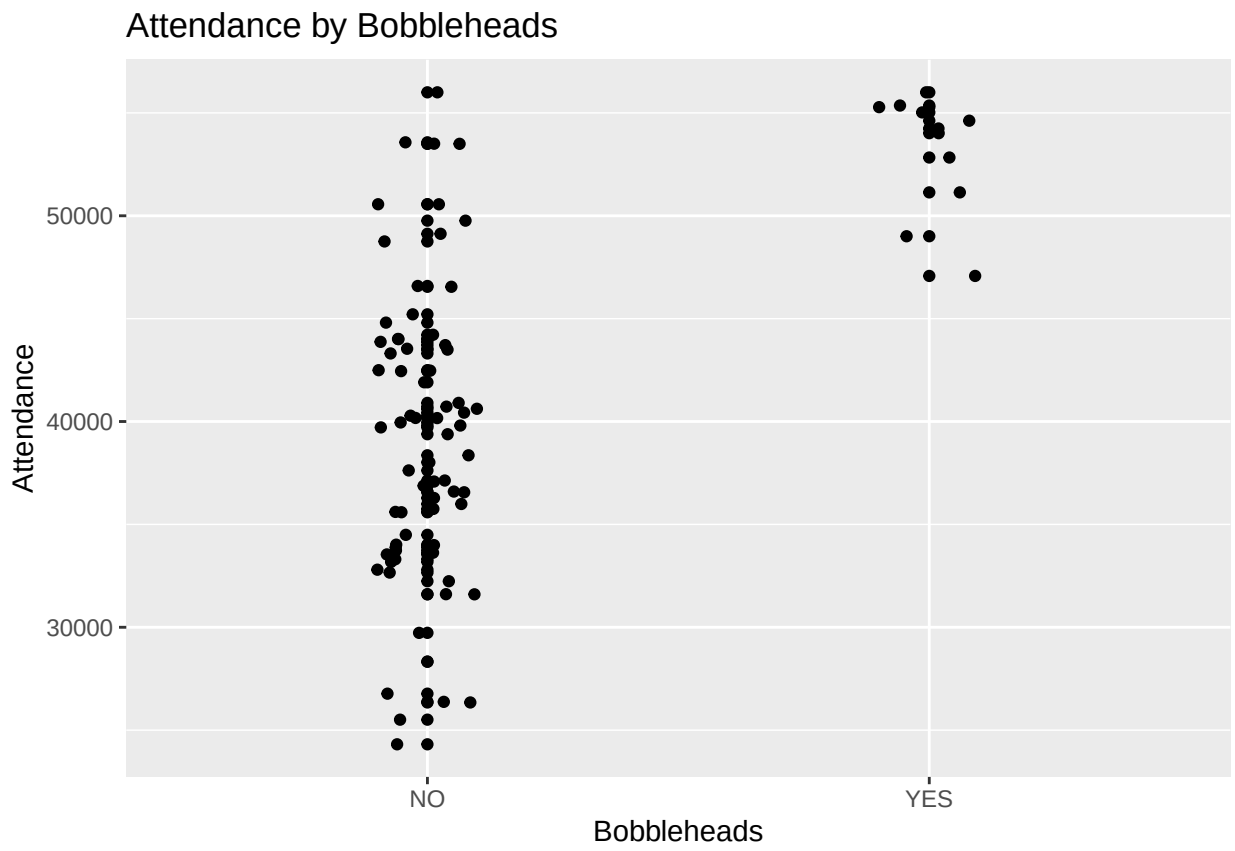

```
# Plot shirts vs attendance
ggplot(df, aes(x=shirt, y=attend)) + geom_point() + geom_jitter(width=0.1) +
  labs(x="Shirt Promotion", y="Attendance", title="Attendance by Shirts")
```




```
# Plot fireworks vs attendance
ggplot(df, aes(x=fireworks, y=attend)) + geom_point() + geom_jitter(width=0.1) +
  labs(x="Fireworks", y="Attendance", title="Attendance by Fireworks")
```




```
# Plot bobbleheads vs attendance
ggplot(df,aes(x=bobblehead, y=attend)) + geom_point() + geom_jitter(width=0.1) +
  labs(x="Bobbleheads", y="Attendance", title="Attendance by Bobbleheads")
```



Shirts and Caps promotions are too few in number to draw conclusions, though shirts may correlate slightly with increase attendance, while fireworks don't seem to have much impact. Bobbleheads on the other hand appear to correlate pretty significantly with attendance. What we don't know is whether people are coming out because they love bobblehead promotions or whether bobbleheads are only distributed on nights with high attendance. This is definitely something to look into since we're interested in increasing attendance.

So the key variables we've honed in on are the month of the year ('month'), the day of the week ('day_of_week'), and whether or not bobbleheads were offered ('bobblehead'). Let's put together a regression with these variables.

```
# Create linear regression model
relation <- lm(formula = attend ~ month + day_of_week + bobblehead, data = df)
summary(relation)
```

```
##
## Call:
## lm(formula = attend ~ month + day_of_week + bobblehead, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10786.5  -3628.1   -516.1   2230.2  14351.0
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          40633.16    2360.11  17.217  < 2e-16 ***
## monthMAY             -2385.62    2291.22  -1.041   0.3015
```

```
## monthJUN              7163.23    2732.72   2.621   0.0108 *
## monthJUL              2849.83    2578.60   1.105   0.2730
## monthAUG              2377.92    2402.91   0.990   0.3259
## monthSEP                29.03    2521.25   0.012   0.9908
## monthOCT              -662.67    4046.45  -0.164   0.8704
## day_of_weekMonday    -6724.00    2506.72  -2.682   0.0092 **
## day_of_weekTuesday    1187.49    2594.66   0.458   0.6487
## day_of_weekWednesday -4263.98    2501.40  -1.705   0.0929 .
## day_of_weekThursday  -5948.64    3339.31  -1.781   0.0794 .
## day_of_weekFriday    -1840.18    2426.79  -0.758   0.4509
## day_of_weekSaturday   -351.95    2417.56  -0.146   0.8847
## bobbleheadYES        10714.90    2419.52   4.429 3.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6120 on 67 degrees of freedom
## Multiple R-squared:  0.5444, Adjusted R-squared:  0.456
## F-statistic: 6.158 on 13 and 67 DF,  p-value: 2.083e-07
```

As expected, May and October are expected to have lower attendance, along with the days Monday and Wednesday, but also Thursday, which I hadn't noticed earlier. Bobbleheads meanwhile appear to have a significant impact on attendance.

Let's use this model to predict attendance.

```
# Set up Data Frame of predictions
noBobble <- data.frame(month=c("MAY", "MAY", "MAY", "OCT", "OCT", "OCT"),
                       day_of_week=c("Monday", "Wednesday", "Thursday"),
                       bobblehead="NO")
# Predict attendance
noBobble$prediction <- predict(relation, noBobble)

# View predictions
noBobble
```

```
##   month day_of_week bobblehead prediction
## 1   MAY      Monday         NO   31523.54
## 2   MAY   Wednesday         NO   33983.56
## 3   MAY    Thursday         NO   32298.90
## 4   OCT      Monday         NO   33246.49
## 5   OCT   Wednesday         NO   35706.52
## 6   OCT    Thursday         NO   34021.86
```

It would appear that Mondays in May (when there are no bobblehead promotions) are expected to have the lowest predicted attendance and should be a good candidate for increasing attendance via promotions.

If in fact it is the bobbleheads that are bringing people out to the games, that might be a good option for these nights. Let's predict the attendance with a bobblehead promotion on Mondays in May.

```
# Create prediction Data Frame
bobble <- data.frame(month="MAY", day_of_week="Monday", bobblehead="YES")

# Predict attendance
bobble$prediction <- predict(relation, bobble)
```

```
# View prediction
bobble
```

```
##   month day_of_week bobblehead prediction
## 1   MAY      Monday        YES   42238.44
```

Adding bobbleheads to the mix accounts for over 10,000 in additional attendance. So, if causitive, this may be the best promotion, assuming the bobbleheads are cost effective. Either way, we should focus on **Mondays in May** for our additional promotional activities.