In [1]:
```r
# Set working directory
setwd("C:\\Users\\myraw\\Jupyter\\DSC630")
```

In [81]:
```r
dodgers <- read.csv(file = 'dodgers.csv')
head(dodgers)
```

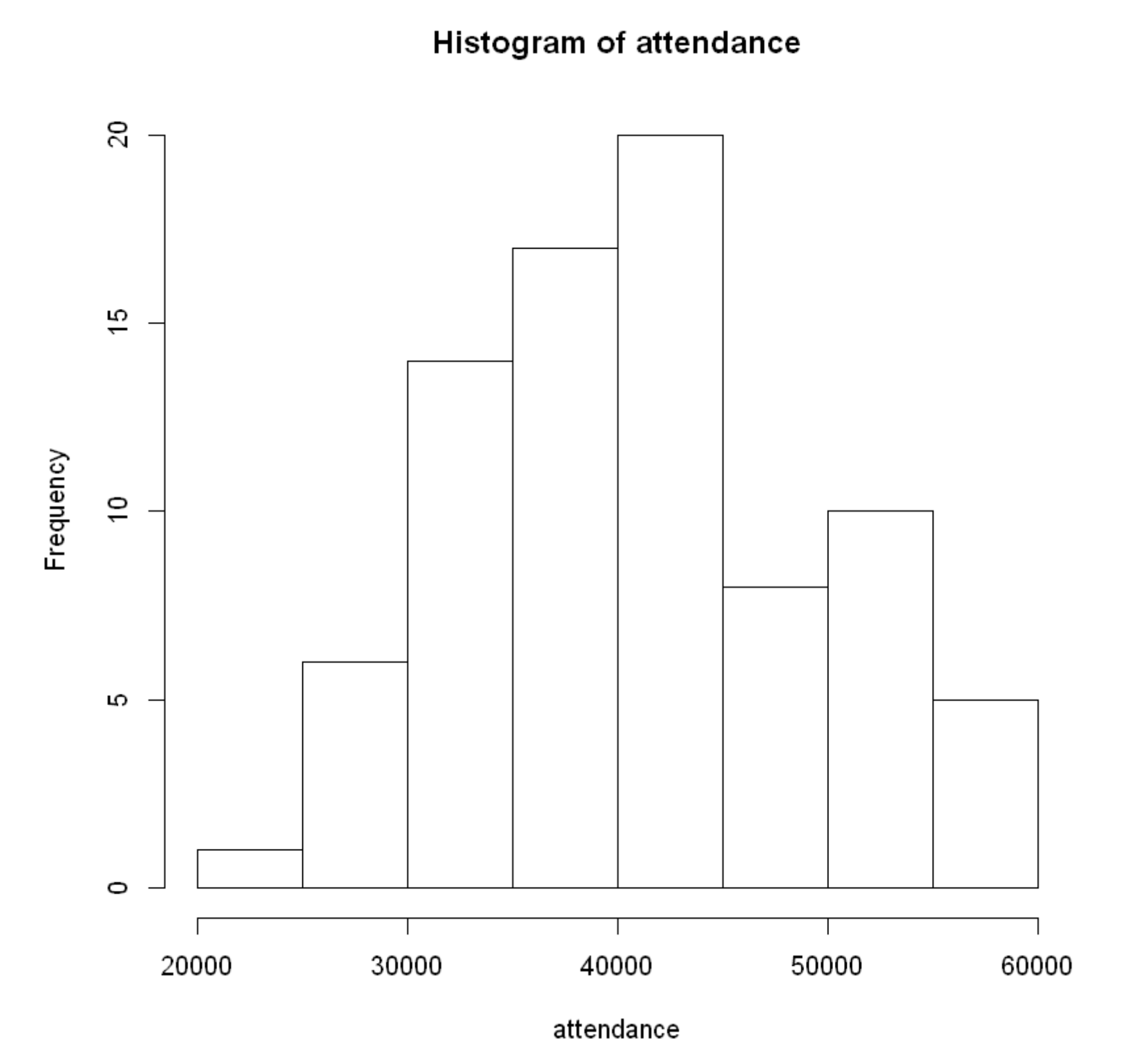
month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks	bobblehead
APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO	NO
APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	NO	NO	NO	NO
APR	12	28328	Thursday	Pirates	57	Cloudy	Night	NO	NO	NO	NO
APR	13	31601	Friday	Padres	54	Cloudy	Night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

In [29]:
```r
# Check the class of all the columns in dataset
lapply(dodgers, class)
```

$month
'factor'
$day
'integer'
$attend
'integer'
$day_of_week
'factor'
$opponent
'factor'
$temp
'integer'
$skies
'factor'
$day_night
'factor'
$cap
'factor'
$shirt
'factor'
$fireworks
'factor'
$bobblehead
'factor'

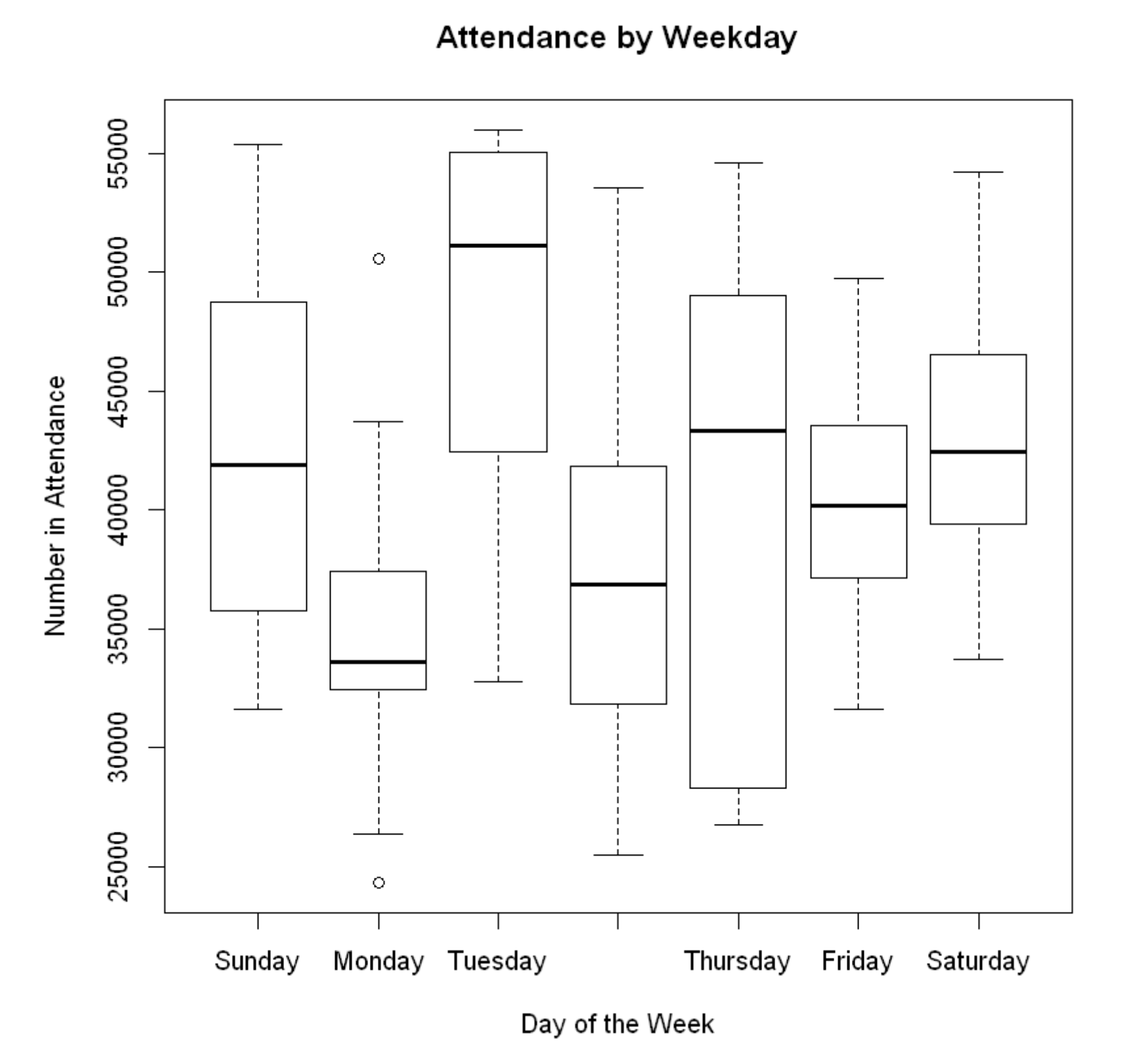
First, I am going to do some exploratory data analysis to determine what approach might be best. I am looking for any stand out deficiencies in attendance rates by month or day of the week to help determine where our marketing promo might have the greatest impact.

In [83]:
```r
# Look at the histogram for attendance
attendance <- dodgers$attend
hist(attendance)
```



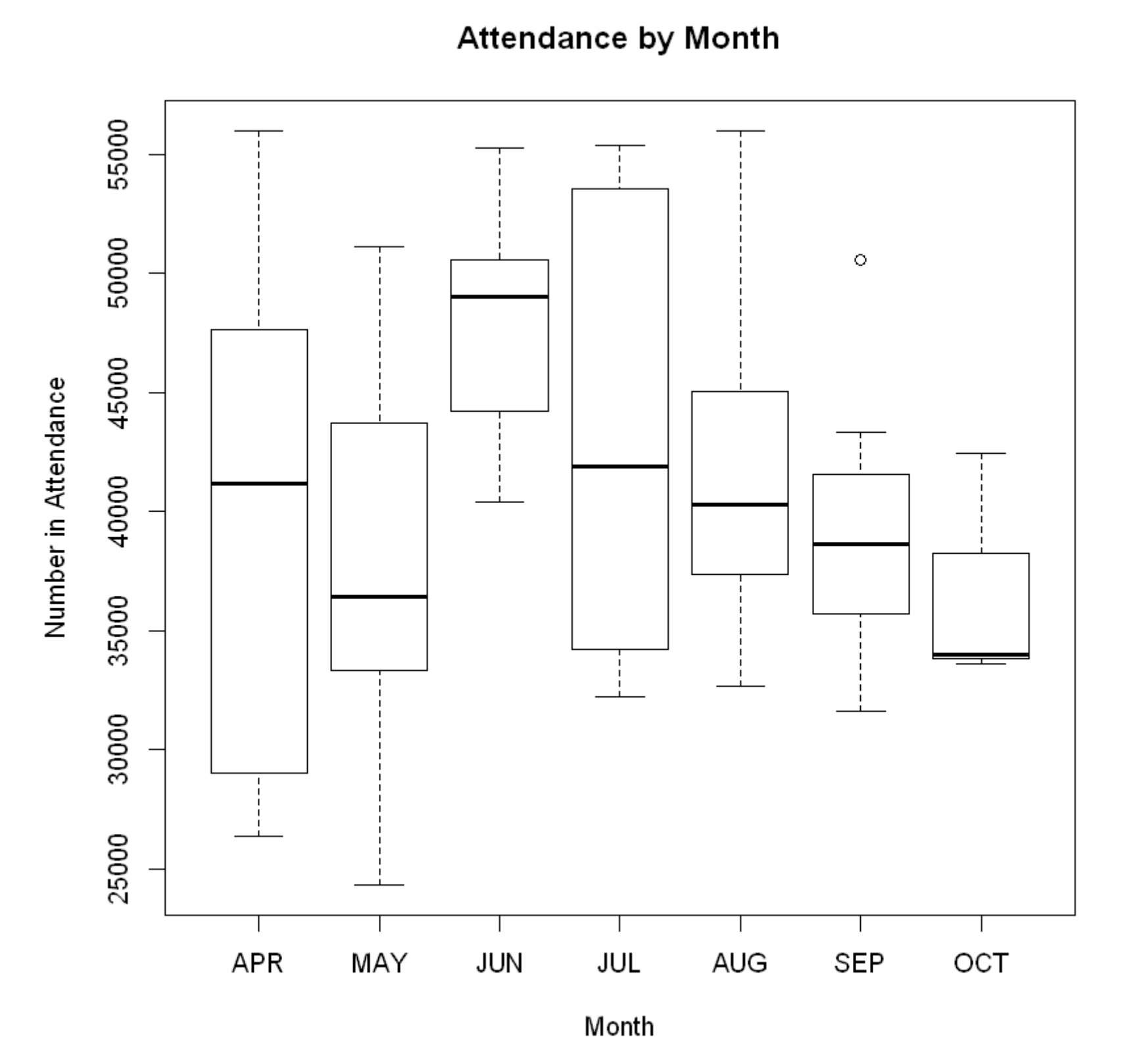
In [2]:
```r
# Reoder days of the week to appear in chronological order versus alphabetical
dodgers$day_of_week <- factor(dodgers$day_of_week , levels=c("Sunday", "Monday", "Tuesday", "Wednesday"
, "Thursday",
                                                            "Friday", "Saturday"))
```

In [14]:
```r
# Look at boxplot for attendance by day of week
boxplot(attend~day_of_week,data=dodgers, main="Attendance by Weekday",
   xlab="Day of the Week", ylab="Number in Attendance")
```



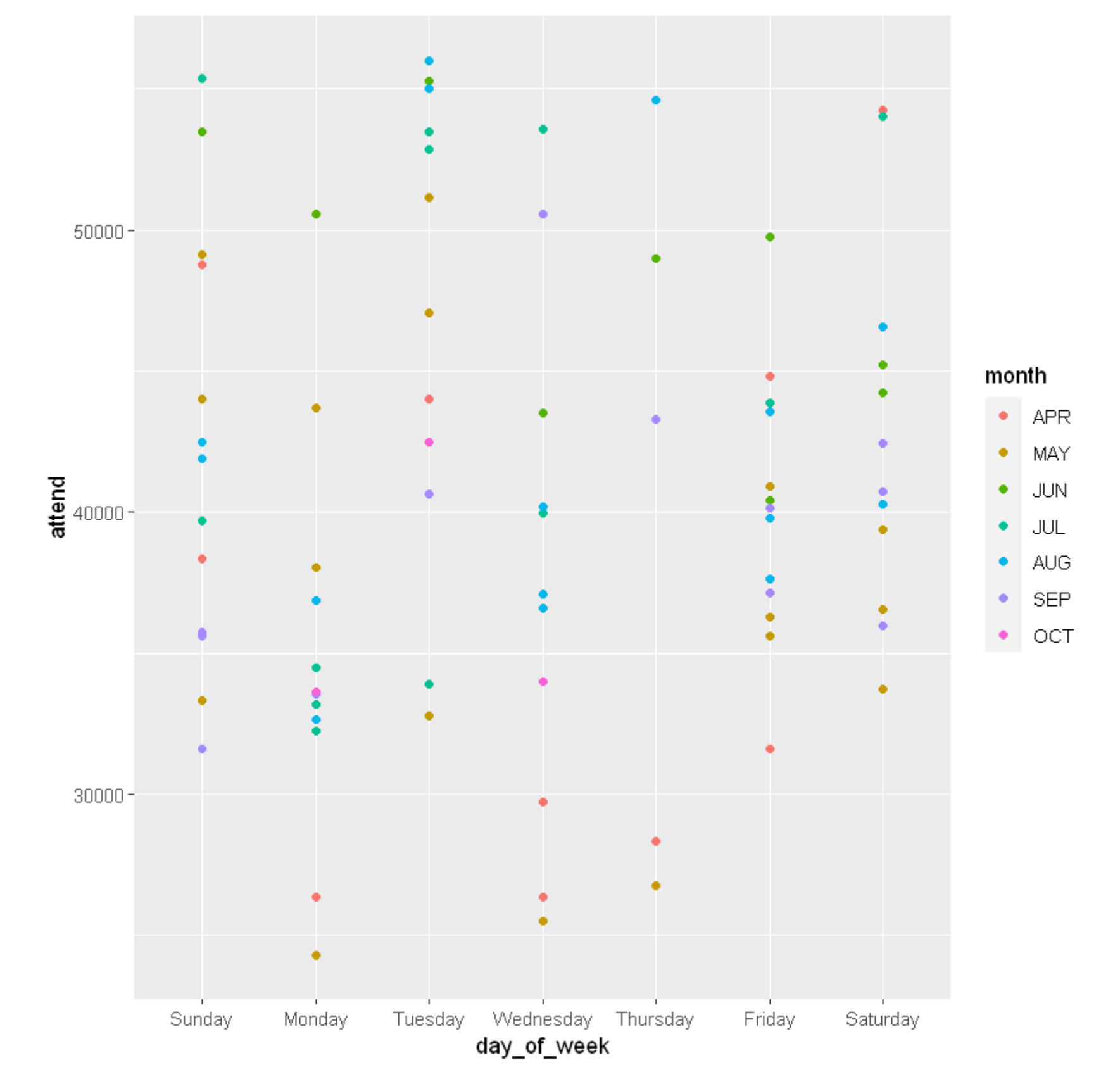
In [3]:
```r
# Reoder months to appear in chronological order versus alphabetical
dodgers$month <- factor(dodgers$month , levels=c("APR", "MAY", "JUN", "JUL", "AUG", "SEP", "OCT"))
```

In [16]:
```r
boxplot(attend~month,data=dodgers, main="Attendance by Month",
   xlab="Month", ylab="Number in Attendance")
```



In [4]:
```r
#Import ggplot library
library(ggplot2)

# Creat scatterplot
ggplot(dodgers, aes(x = day_of_week, y = attend)) +
  geom_point(aes(color = month))
```

```
Warning message:
"package 'ggplot2' was built under R version 3.6.3"
```



Mondays and Wednesdays definitely stick out as far as low attendance goes. A couple months don't look so hot (May and Oct), but when I plotted attendance by day of week and month, it appears that Mondays, Wednesdays, and Thursdays in May and April generally have the worst attendance (less that 30000 per game).

Before creating a model, I want to use RFE to find out what variables have the most influence over attendance.

In [35]:
```r
# Install packages to use RFE
library("dplyr")
library("faux")
library("DataExplorer")
library("caret")
library("randomForest")
```

In [36]:
```r
# Define the control using a random forest selection function
control <- rfeControl(functions = rfFuncs, # random forest
                      method = "repeatedcv", # repeated cv
                      repeats = 5, # number of repeats
                      number = 10) # number of folds
```

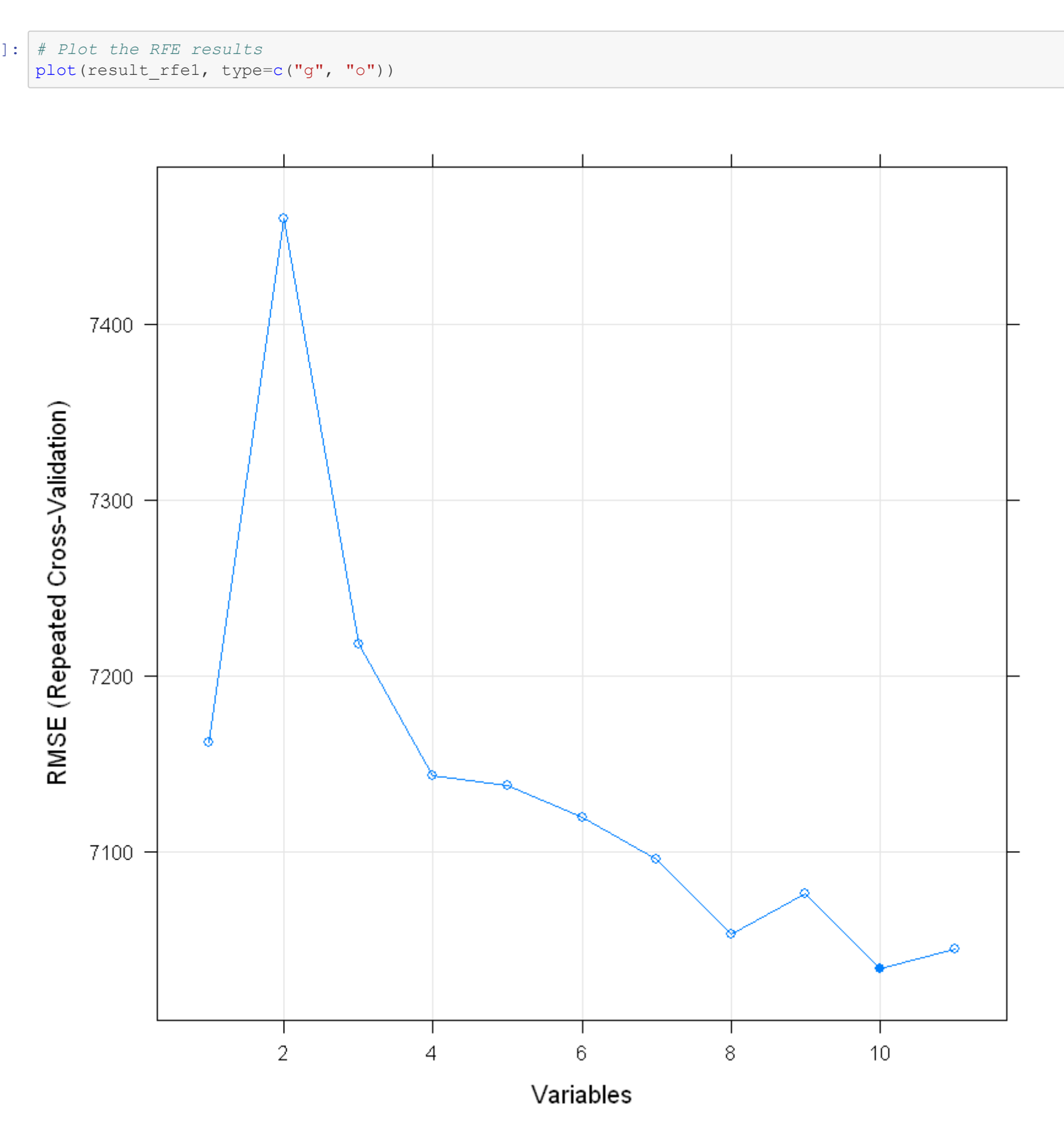
In [42]:
```r
# Split data into train-test sets

# Features
x <- dodgers %>%
  select(-attend) %>%
  as.data.frame()

# Target variable
y <- dodgers$attend

# Training: 80%; Test: 20%
set.seed(2021)
inTrain <- createDataPartition(y, p = .80, list = FALSE)[,1]

x_train <- x[ inTrain, ]
x_test  <- x[-inTrain, ]

y_train <- y[ inTrain]
y_test  <- y[-inTrain]
```

In [43]:
```r
# Run RFE
result_rfe1 <- rfe(x = x_train,
                   y = y_train,
                   sizes = c(1:11),
                   rfeControl = control)

# Print the results
result_rfe1

# Print the selected features
predictors(result_rfe1)
```

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance over subset size:

 Variables RMSE Rsquared  MAE RMSESD RsquaredSD  MAESD Selected
         1 7162   0.4683 5815   1544     0.2216 1335.3
         2 7460   0.3056 5961   1358     0.2434 1254.2
         3 7218   0.3511 5782   1323     0.2718 1280.1
         4 7143   0.3648 5822   1130     0.2740 1008.0
         5 7137   0.3745 5850   1131     0.2780 1046.0
         6 7120   0.3430 5850   1094     0.2693  881.5
         7 7096   0.3573 5859   1108     0.2779  931.0
         8 7053   0.3637 5817   1053     0.2727  851.2
         9 7076   0.3573 5837   1071     0.2652  871.7
        10 7034   0.3658 5799   1066     0.2735  871.3        *
        11 7045   0.3640 5825   1030     0.2720  868.9

The top 5 variables (out of 10):
   bobblehead, day_of_week, opponent, month, temp
```

'bobblehead' 'day_of_week' 'opponent' 'month' 'temp' 'shirt' 'fireworks' 'day' 'cap' 'day_night'

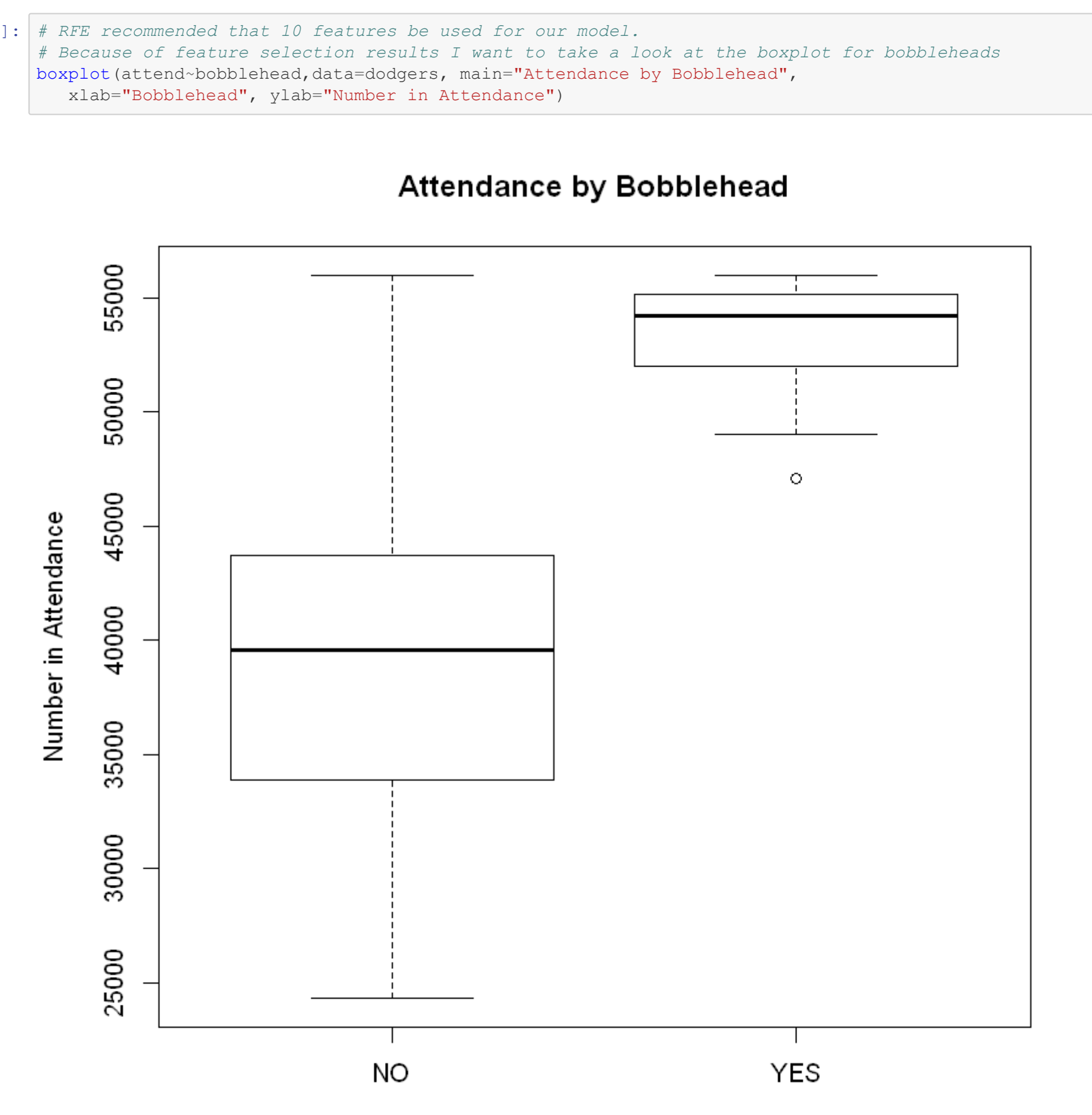
In [80]:
```r
# Plot the RFE results
plot(result_rfe1, type=c("g", "o"))
```



In [73]:
```r
# Use the model on test data
postResample(predict(result_rfe1, x_test), y_test)
```

RMSE	7428.48765690282
Rsquared	0.183578718325904
MAE	5940.22107138276

In [49]:
```r
# RFE recommended that 10 features be used for our model.
# Because of feature selection results I want to take a look at the boxplot for bobbleheads
boxplot(attend~bobblehead,data=dodgers, main="Attendance by Bobblehead",
   xlab="Bobblehead", ylab="Number in Attendance")
```



In [59]:
```r
# Now I am going to create a few models and use the anova test

# create baseline model for attendance. Use poisson regression model because attend is count data.
model_base <- glm(attend ~ 1, family = "poisson", data=dodgers)

# create model for attendance by days of week.
model_days <- glm(attend ~ day_of_week, family = "poisson", data=dodgers)

# create model for attendance by days of week and month
model_both <- glm(attend ~ day_of_week + month, family = "poisson", data=dodgers)

# create modelrecommended by RFE
model_rfe <- glm(attend ~ day_of_week + opponent + month + bobblehead + temp + shirt + fireworks + day
+ cap + day_night,
                 family = "poisson", data=dodgers)
```

In [74]:
```r
# Compare models
anova(model_base, model_days, model_both, model_rfe)
```

Resid. Df	Resid. Dev	Df	Deviance
80	135065.17	NA	NA
74	104517.80	6	30547.37
68	80314.34	6	24203.46
45	41725.39	23	38588.95

In [69]:

```
# The RFE recommended model had lower residual deviance, meaning it was a better fit and an R2 of 0.68,
 which is substantial.
# Let's take a look at the summary for it.
summary(model_rfe)
```

```
Call:
glm(formula = attend ~ day_of_week + opponent + month + bobblehead +
    temp + shirt + fireworks + day + cap + day_night, family = "poisson",
    data = dodgers)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-46.441  -17.190   -0.601    7.773   62.856

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          10.7196804  0.0138345 774.852  < 2e-16 ***
day_of_weekMonday    -0.0179428  0.0040756  -4.403 1.07e-05 ***
day_of_weekTuesday    0.2009465  0.0039921  50.336  < 2e-16 ***
day_of_weekWednesday -0.0126775  0.0038755  -3.271  0.00107 **
day_of_weekThursday   0.0076039  0.0046081   1.650  0.09891 .
day_of_weekFriday    -0.4340108  0.0073674 -58.910  < 2e-16 ***
day_of_weekSaturday   0.0729555  0.0031977  22.815  < 2e-16 ***
opponentAstros       -0.5187919  0.0114120 -45.460  < 2e-16 ***
opponentBraves       -0.5008576  0.0111279 -45.009  < 2e-16 ***
opponentBrewers      -0.5894965  0.0121561 -48.494  < 2e-16 ***
opponentCardinals    -0.3166021  0.0103432 -30.610  < 2e-16 ***
opponentCubs         -0.2573975  0.0099225 -25.941  < 2e-16 ***
opponentGiants       -0.4343812  0.0105113 -41.325  < 2e-16 ***
opponentMarlins      -0.4684950  0.0110484 -42.404  < 2e-16 ***
opponentMets         -0.1031137  0.0049801 -20.705  < 2e-16 ***
opponentNationals    -0.1855830  0.0109833 -16.897  < 2e-16 ***
opponentPadres       -0.2735250  0.0088208 -31.009  < 2e-16 ***
opponentPhillies     -0.3348028  0.0100148 -33.431  < 2e-16 ***
opponentPirates      -0.3250357  0.0107596 -30.209  < 2e-16 ***
opponentReds         -0.4044556  0.0095943 -42.156  < 2e-16 ***
opponentRockies      -0.4266107  0.0101985 -41.831  < 2e-16 ***
opponentSnakes       -0.4975249  0.0101226 -49.150  < 2e-16 ***
opponentWhite Sox    -0.0043519  0.0044275  -0.983  0.32564
monthMAY              0.0997543  0.0051015  19.554  < 2e-16 ***
monthJUN             -0.1153495  0.0095664 -12.058  < 2e-16 ***
monthJUL              0.0811355  0.0051108  15.875  < 2e-16 ***
monthAUG              0.1766884  0.0062892  28.094  < 2e-16 ***
monthSEP              0.0546010  0.0064094   8.519  < 2e-16 ***
monthOCT              0.1152253  0.0080424  14.327  < 2e-16 ***
bobbleheadYES         0.1923710  0.0025538  75.328  < 2e-16 ***
temp                  0.0014737  0.0002031   7.257 3.95e-13 ***
shirtYES              0.0331600  0.0035912   9.234  < 2e-16 ***
fireworksYES          0.4695285  0.0066310  70.809  < 2e-16 ***
day                   0.0037936  0.0001188  31.929  < 2e-16 ***
capYES               -0.1571928  0.0049314 -31.876  < 2e-16 ***
day_nightNight       -0.0828032  0.0029574 -27.999  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 135065  on 80  degrees of freedom
Residual deviance:  41725  on 45  degrees of freedom
AIC: 42805

Number of Fisher Scoring iterations: 4
```

Upon conducting analysis of the data my recommendation would be to run the marketing promo on Mondays in October. May has less attendees on Mondays in total, but October presents a slightly larger capacity for impact on attendance numbers. For the promotion, fireworks is the leading promotion, should that not be available, bobbleheads make the second most impact to total number of attendees.