Spring 2023
GE 461 Introduction to Data Science - Project 1

Contents

Project 1

T he Dodgers is a professional baseball team and plays in the Major Baseball League. The team owns a 56,000-seat stadium and is interested in increasing the attendance of their fans during home games. *At the moment the team management would like to know if bobblehead promotions increase the attendance of the team's fans?*

The 2012 season data in the `events` table of SQLite database `data/dodgers.sqlite` contain for each of 81 home play the

- month,

- day,
- weekday,
- part of the day (day or night),
- attendance,
- opponent,
- temperature,
- whether cap or shirt or bobblehead promotions were run, and
- whether fireworks were present.

Download the Dataset

Connect to `data/dodgers.sqlite`. Read table `events` into a variable in `R`.

```
library(RSQLite)
con <- dbConnect(SQLite(), "C:/Users/yagiz/Desktop/4-2/GE-461/PromotionAnalysis/data/dodgers.sqlite")

events <- dbReadTable(con, "events")
events <- as.data.table(events)

rm(con)
```

Some Manipulations

```
events[, day_of_week := factor(day_of_week, levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "F

events[, month := factor(month, levels = c("APR","MAY","JUN","JUL","AUG","SEP","OCT"))]

events[, lapply(.SD, function(x) if(is.character(x)) factor(x) else x)]
```

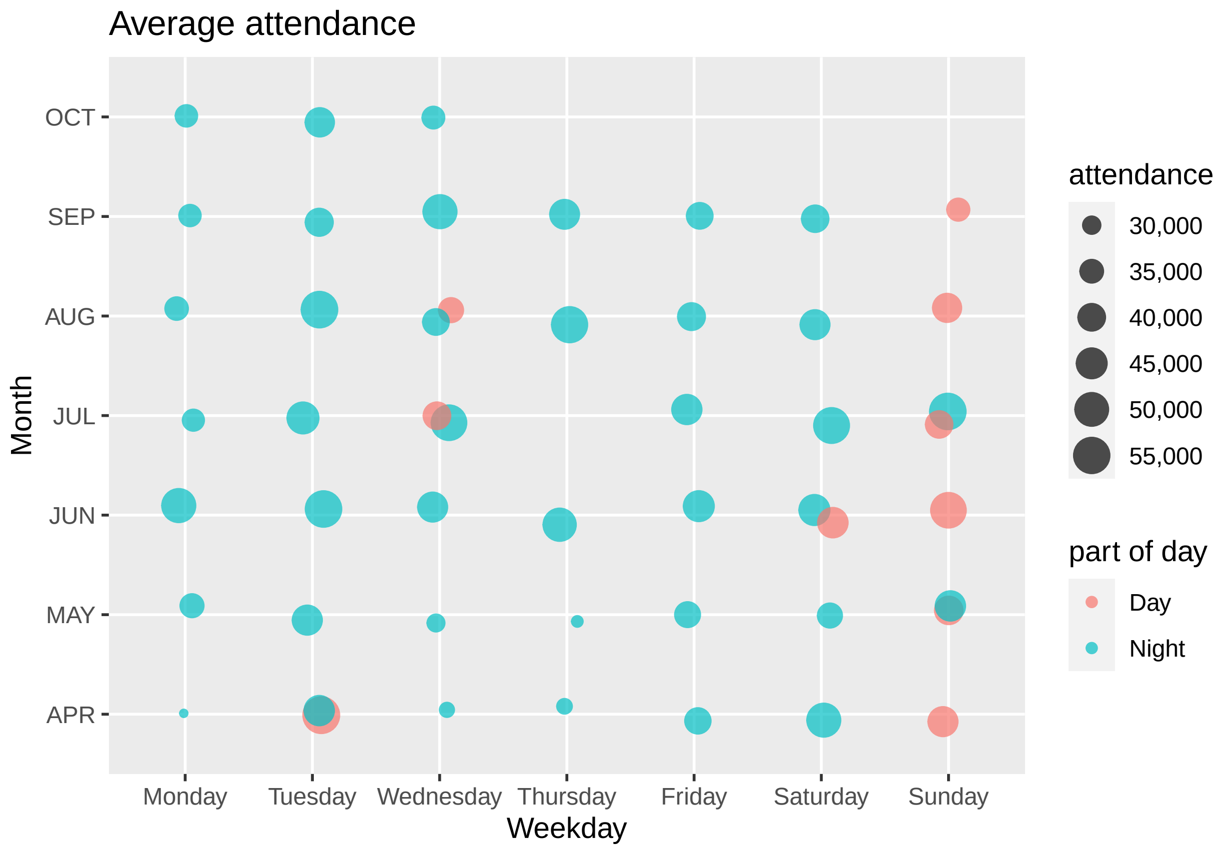
```
events[, temp := round((temp- 32)*5/9)] # convert the temperature variable from Fahrenheit to Celsius
```

Preliminary Analysis

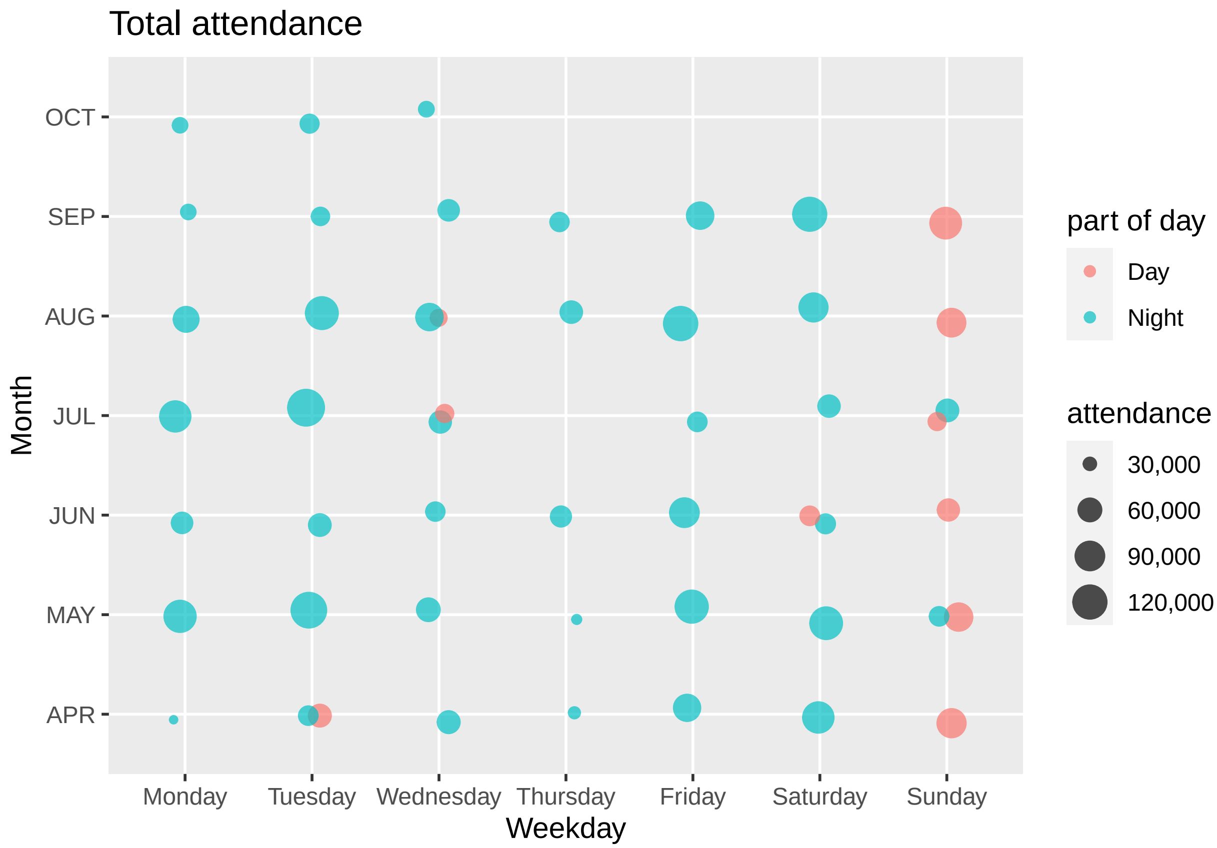
```
events[, .(total_attend = sum(attend)), month][order(-total_attend)]
```

```
sum_attend <- events[, .(mean_attend = mean(attend),
                         total_attend = sum(attend)),
                         by = .(day_of_week, month, day_night)]
```

```
ggplot(data=sum_attend,aes(day_of_week, month, month)) +
geom_jitter(aes(size = mean_attend, col = day_night), width = .1, height = .1, alpha=0.7) +
scale_size(labels = scales::comma) +
labs(title = "Average attendance", size = "attendance", col = "part of day",
     x = "Weekday", y = "Month")
```




```r
ggplot(data=sum_attend, aes(day_of_week, month)) +
geom_jitter(aes(size = total_attend, col = day_night), width = .1, height = .1, alpha=0.7) +
labs(title = "Total attendance", size = "attendance", col = "part of day",
     x = "Weekday", y = "Month") +
scale_size(labels = scales::comma) +
guides(col = guide_legend(order = 1),
       shape = guide_legend(order = 2))
```



From the above two graphs, it can be observed that the games that are played in **day** are generally on Sunday and most of the games are played at night. There aren't many games played in October and average attendance is relatively low in May.

There are 4 types of promotions which are cap, shirt, fireworks and bobblehead.

```
## The number of occurrences of matches with no promotions is 51 .
## The number of matches with only bobblehead promotion is 11 .
## The number of matches with only cap promotion is  2 .
## The number of matches with only shirt promotion is  3 .
## The number of matches with only firework promotion is  14
```

When we sum all of these, 51+11+2+3+14=81, which is the whole dataset. So, it is observed that there is at most one type of advertising in each match. For example, if bobblehead is 'YES' than the rest of the promotions are 'NO'.

Number of occurrences ('YES') for each type of advertising is as follows:

```
## cap = 2
## shirt = 3
## fireworks = 14
## bobblehead = 11
```

Cap and shirt promotions are applied rarerly.

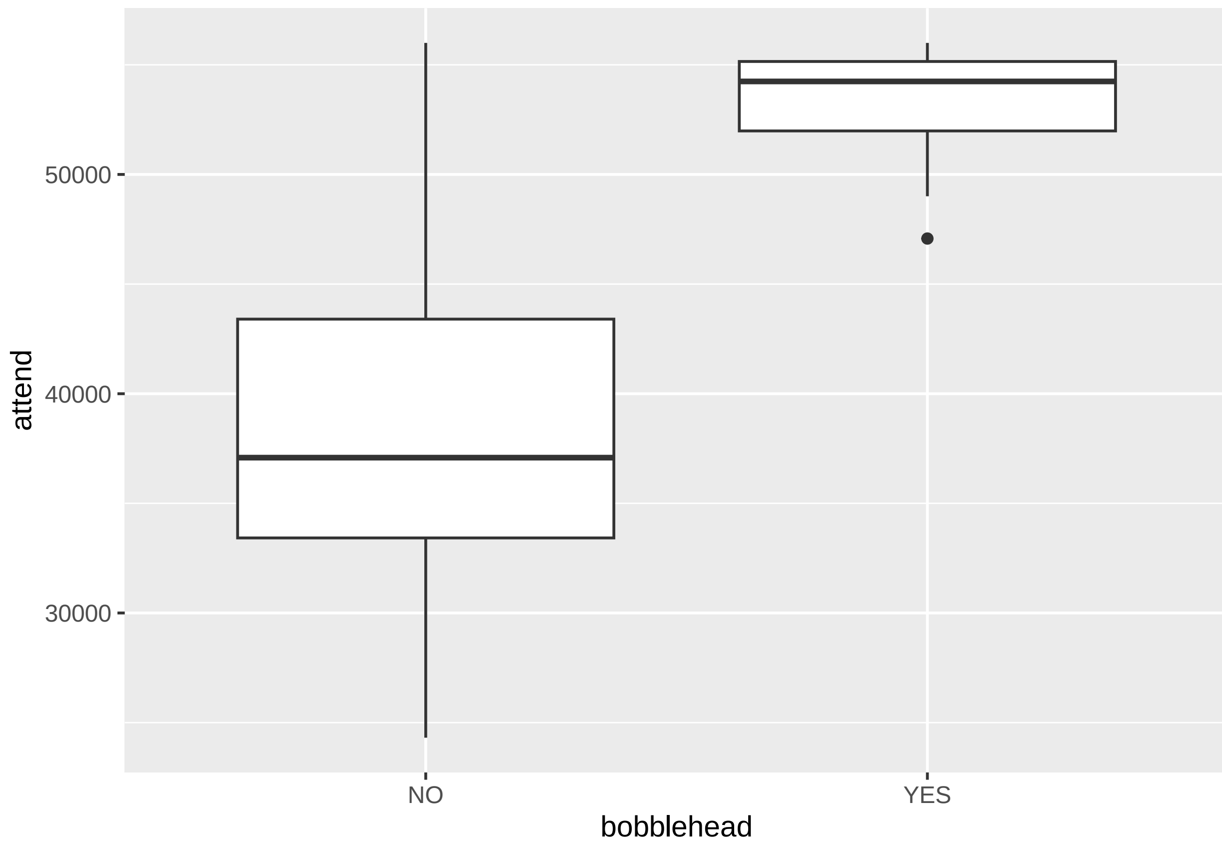
Variable Exploration

booblehead

Checking the effect of a promotion using the whole dataset does not seem right. It is possible that the shirt promotion increases the number of fans but because there are also other promotions in other matches where the shirt promotion does not exist, it would not be possible to see the increasing effect of the shirt promotion. The reason is the other promotions may also increase the number of fans. So, we should compare the effect of each promotion in regular days that do not have any promotion.

Now we will answer the question of "does the bobblehead promotion have a statistically significant effect on the attendance?".

```
ggplot(data=events[cap=="NO" & shirt=="NO" & fireworks=="NO"], aes(bobblehead, attend)) +
geom_boxplot()
```



From the boxplot it is observed that having bobblehead strictly increases the attendance. The median without bobblehead is around 37500, but it is around 58000 with bobblehead. They do not share any observation in their IQR.

We explored a relationship between bobblehead and attendance, but we should be able to statistically explain this relationship.

```
t.test(events[cap=="NO" & shirt=="NO" & fireworks=="NO" & bobblehead=="YES", attend],
       events[cap=="NO" & shirt=="NO" & fireworks=="NO" & bobblehead=="NO", attend])
```

```
##
##  Welch Two Sample t-test
##
```

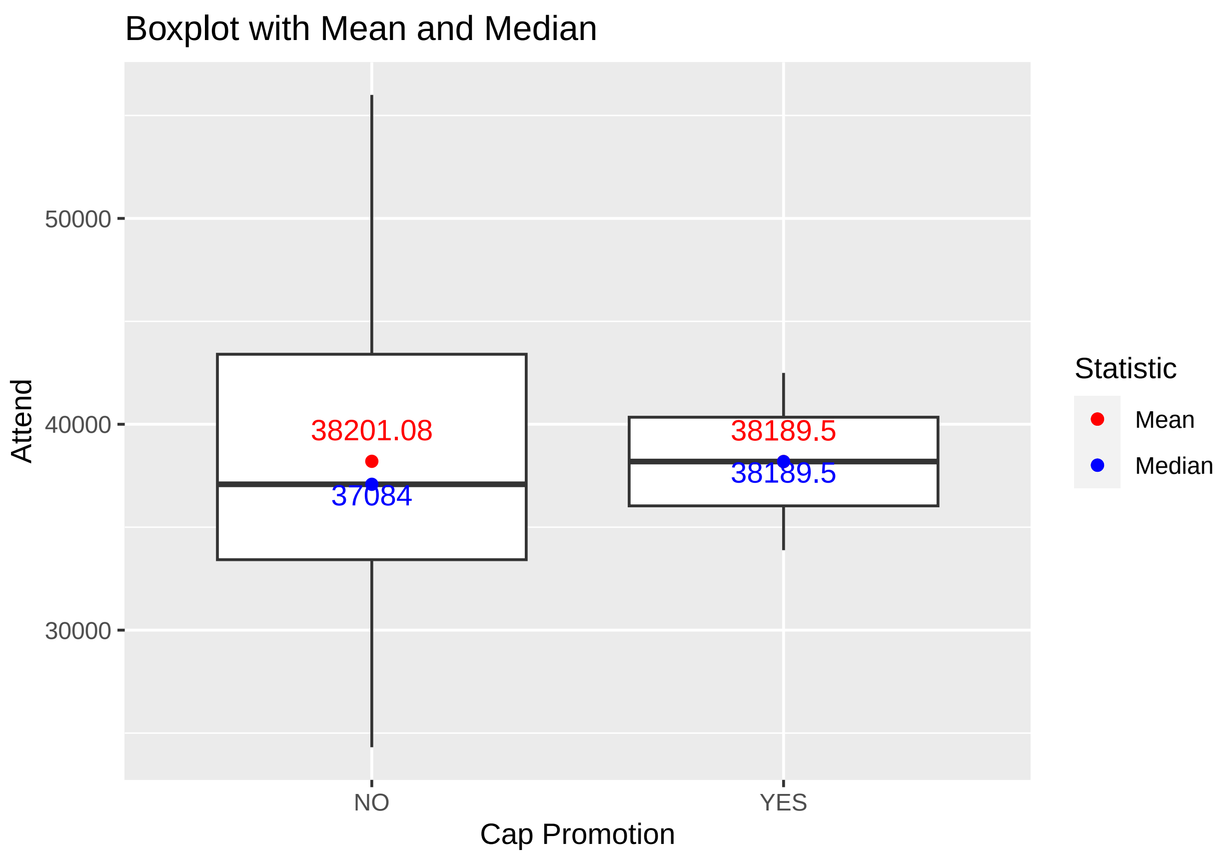
```
## data:  events[cap == "NO" & shirt == "NO" & fireworks == "NO" & bobblehead == "YES", attend] and ever
## t = 11.01, df = 41.93, p-value = 0.00000000000006013
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12204.22 17682.89
## sample estimates:
## mean of x mean of y
##  53144.64  38201.08
```

The statistical test indicates that there is a significant difference between the two groups. So, indeed bobblehead has a statistically significant effect on attendance.

cap

Start by checking the boxplot.

```
ggplot(data=events[bobblehead=="NO" & shirt=="NO" & fireworks=="NO"], aes(cap, attend)) +
geom_boxplot() +
stat_summary(fun = mean, geom = "point", aes(color = "Mean"),
             shape = 16, size = 2) +
stat_summary(fun = median, geom = "point", aes(color = "Median"),
             shape = 16, size = 2) +
stat_summary(fun = mean, geom = "text", vjust = -1,
             aes(label = round(..y.., 2)), color = "red") +
stat_summary(fun = median, geom = "text", vjust = 1,
             aes(label = round(..y.., 2)), color = "blue") +
scale_color_manual(name = "Statistic",
                   values = c("Mean" = "red", "Median" = "blue")) +
labs(x = "Cap Promotion", y = "Attend", title = "Boxplot with Mean and Median")
```



The plot shows that the mean and median attend under with cap and without cap do not differ significantly.

```
t.test(events[bobblehead=="NO" & shirt=="NO" & fireworks=="NO"&cap=="YES", attend],
       events[bobblehead=="NO" & shirt=="NO" & fireworks=="NO"&cap=="NO", attend])
```

```
##
##  Welch Two Sample t-test
##
## data:  events[bobblehead == "NO" & shirt == "NO" & fireworks == "NO" & cap == "YES", attend] and even
## t = -0.0026138, df = 1.1204, p-value = 0.9983
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -43838.39  43815.24
## sample estimates:
## mean of x mean of y
##  38189.50  38201.08
```

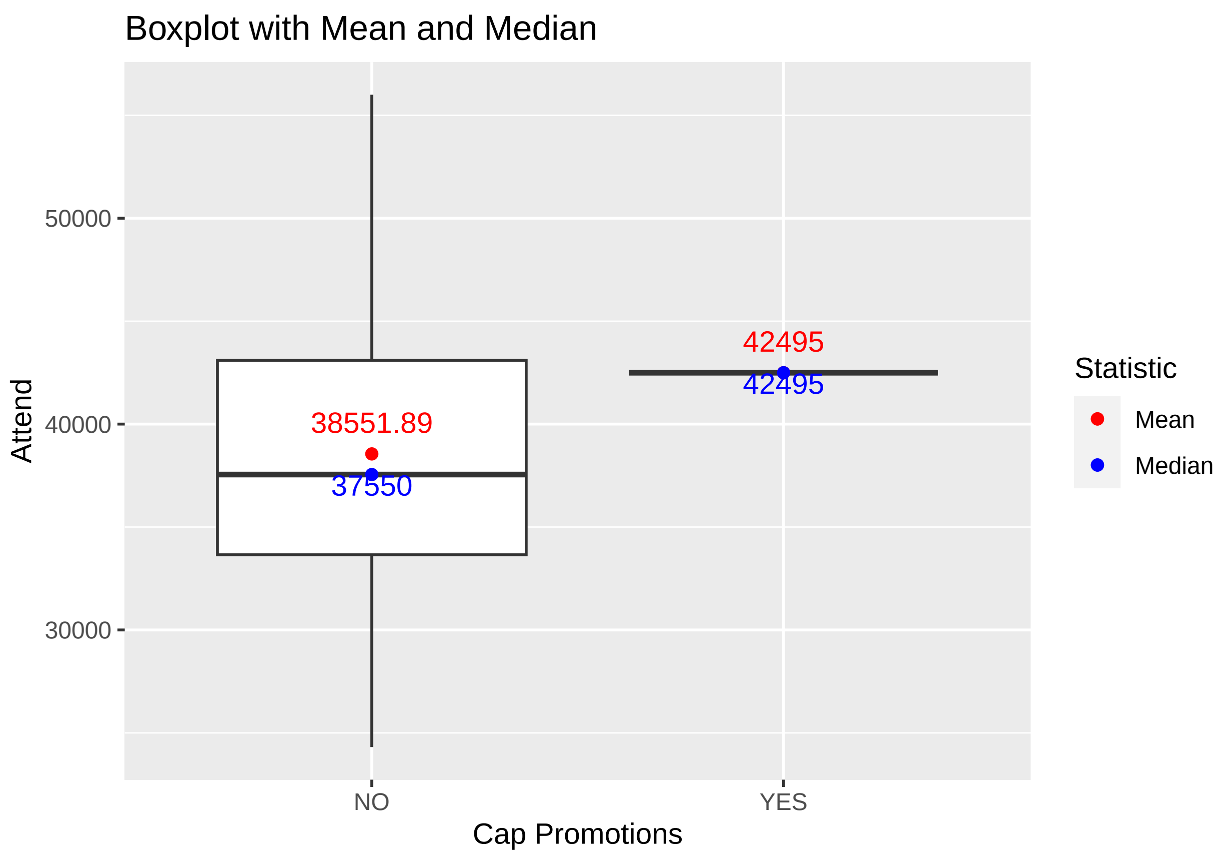
The t-test supports our argument. So, there is no relationship between cap and attendance.

cap & skies

Also, because there are not different promotions applied on same day, it is not possible to check combinations of promotions. However, it is possible that in clear days, the cap promotion increases the attendance because people may want to wear the cap to protect their heads from the sun. Now we are going to analyze the boxplot again but only in clear days.

```
  ggplot(data=events[bobblehead=="NO" & shirt=="NO" & fireworks=="NO" & skies == "Clear"], aes(cap, att
  geom_boxplot() +
```

```r
stat_summary(fun = mean, geom = "point", aes(color = "Mean"),
             shape = 16, size = 2) +
stat_summary(fun = median, geom = "point", aes(color = "Median"),
             shape = 16, size = 2) +
stat_summary(fun = mean, geom = "text", vjust = -1,
             aes(label = round(..y.., 2)), color = "red") +
stat_summary(fun = median, geom = "text", vjust = 1,
             aes(label = round(..y.., 2)), color = "blue") +
scale_color_manual(name = "Statistic",
                   values = c("Mean" = "red", "Median" = "blue")) +
labs(x = "Cap Promotions", y = "Attend", title = "Boxplot with Mean and Median")
```



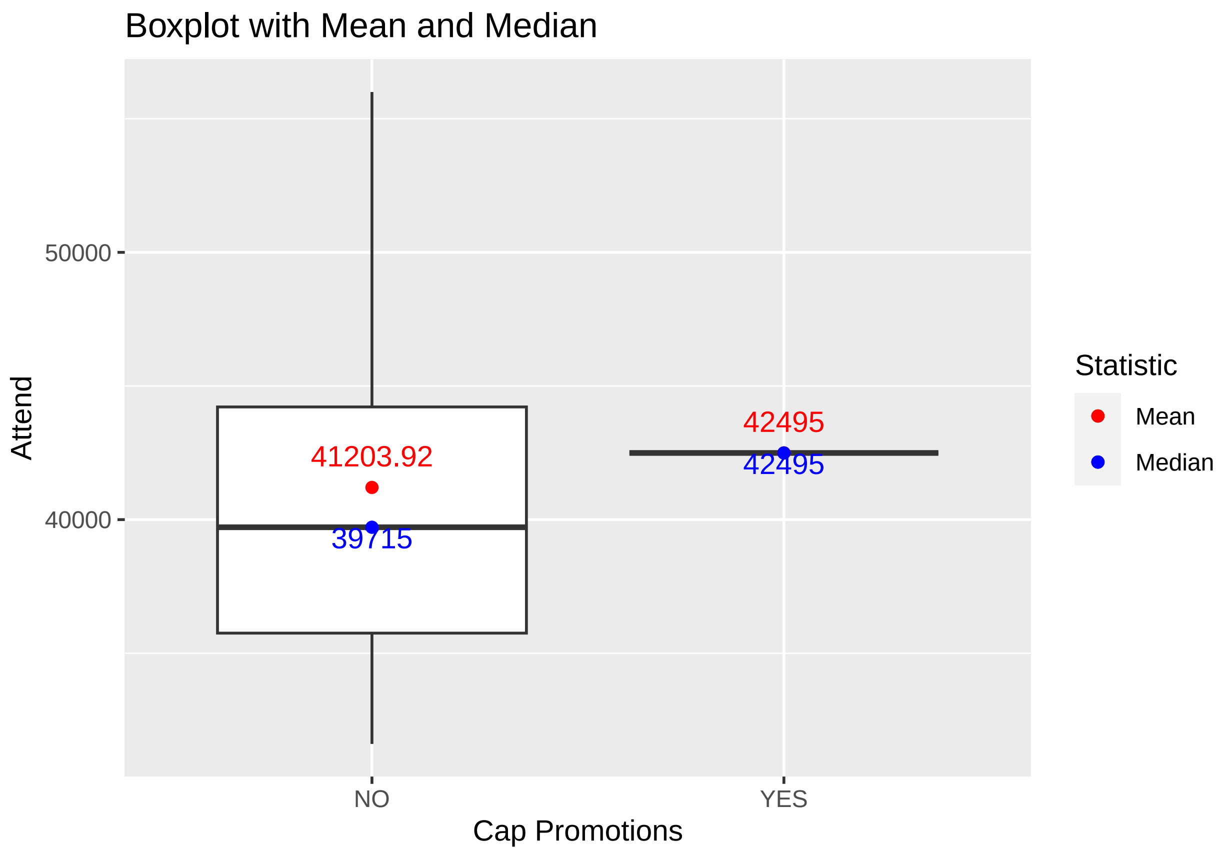
There is only one day where cap == "YES" & skies=="clear". So, it seems, we cannot test our hypothesis.

cap & day_night

Also, it is reasonable to think that the cap promotion may have more affect when it is day rather than night.

```r
ggplot(data=events[bobblehead=="NO" & shirt=="NO" & fireworks=="NO" & day_night == "Day"], aes(cap, a
geom_boxplot() +
stat_summary(fun = mean, geom = "point", aes(color = "Mean"),
             shape = 16, size = 2) +
stat_summary(fun = median, geom = "point", aes(color = "Median"),
             shape = 16, size = 2) +
stat_summary(fun = mean, geom = "text", vjust = -1,
             aes(label = round(..y.., 2)), color = "red") +
stat_summary(fun = median, geom = "text", vjust = 1,
```

```
              aes(label = round(..y.., 2)), color = "blue") +
scale_color_manual(name = "Statistic",
                   values = c("Mean" = "red", "Median" = "blue")) +
labs(x = "Cap Promotions", y = "Attend", title = "Boxplot with Mean and Median")
```

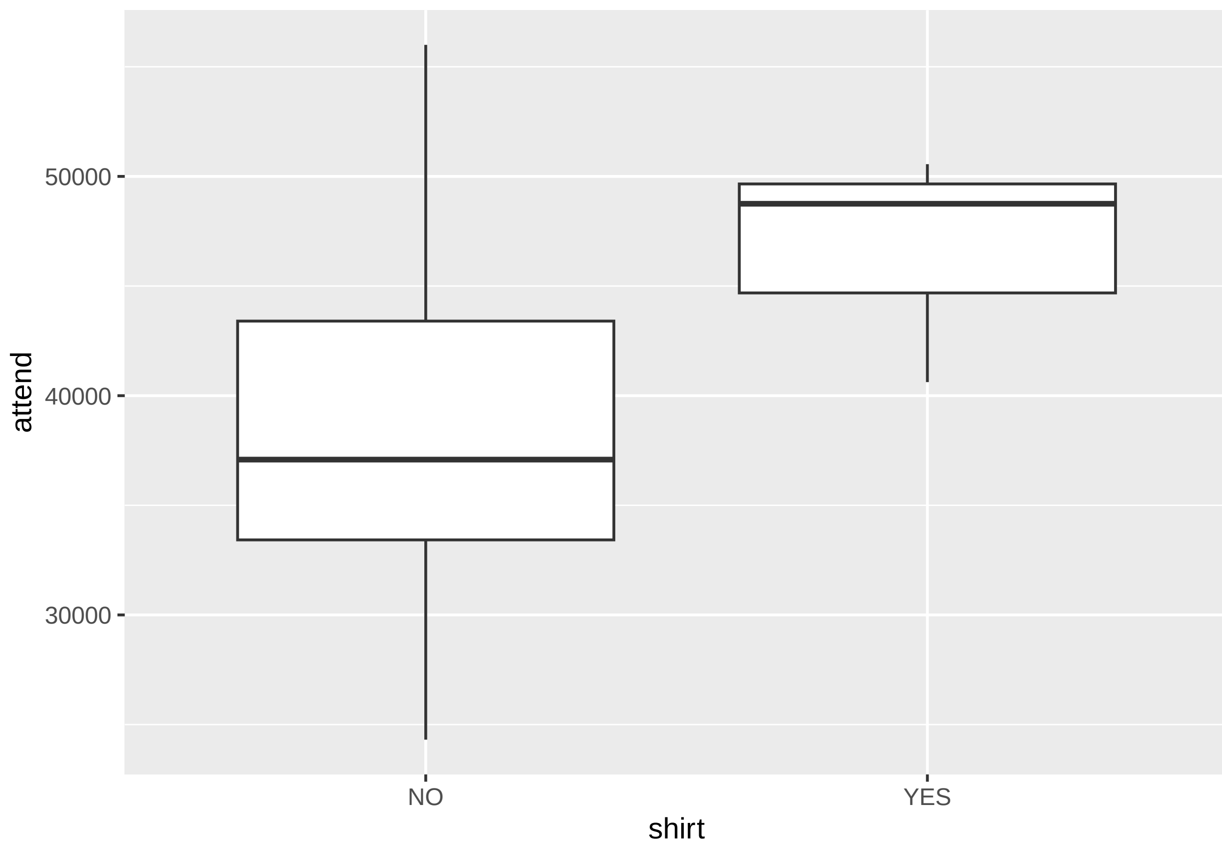


Again we do have only 1 observation, so it is not possible again to test this hypothesis.

shirt

Let's check the effect of the shirt promotion.

```
ggplot(data=events[bobblehead=="NO" & cap=="NO" & fireworks=="NO"], aes(shirt, attend)) +
geom_boxplot()
```



The boxplot suggests that the shirt promotion has a significant impact on the attendance.

```
t.test(events[bobblehead=="NO" & cap=="NO" & fireworks=="NO"& shirt=="YES", attend],
       events[bobblehead=="NO" & cap=="NO" & fireworks=="NO"& shirt=="NO", attend])
```

```
##
##  Welch Two Sample t-test
##
## data:  events[bobblehead == "NO" & cap == "NO" & fireworks == "NO" & shirt == "YES", attend] and even
## t = 2.6141, df = 2.4898, p-value = 0.09626
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3136.222 20021.399
## sample estimates:
## mean of x mean of y
##  46643.67  38201.08
```

The p-value is 0.09626. So, we do not see a statistically significant difference between the average attendance of the games played under the shirt promotion or not.

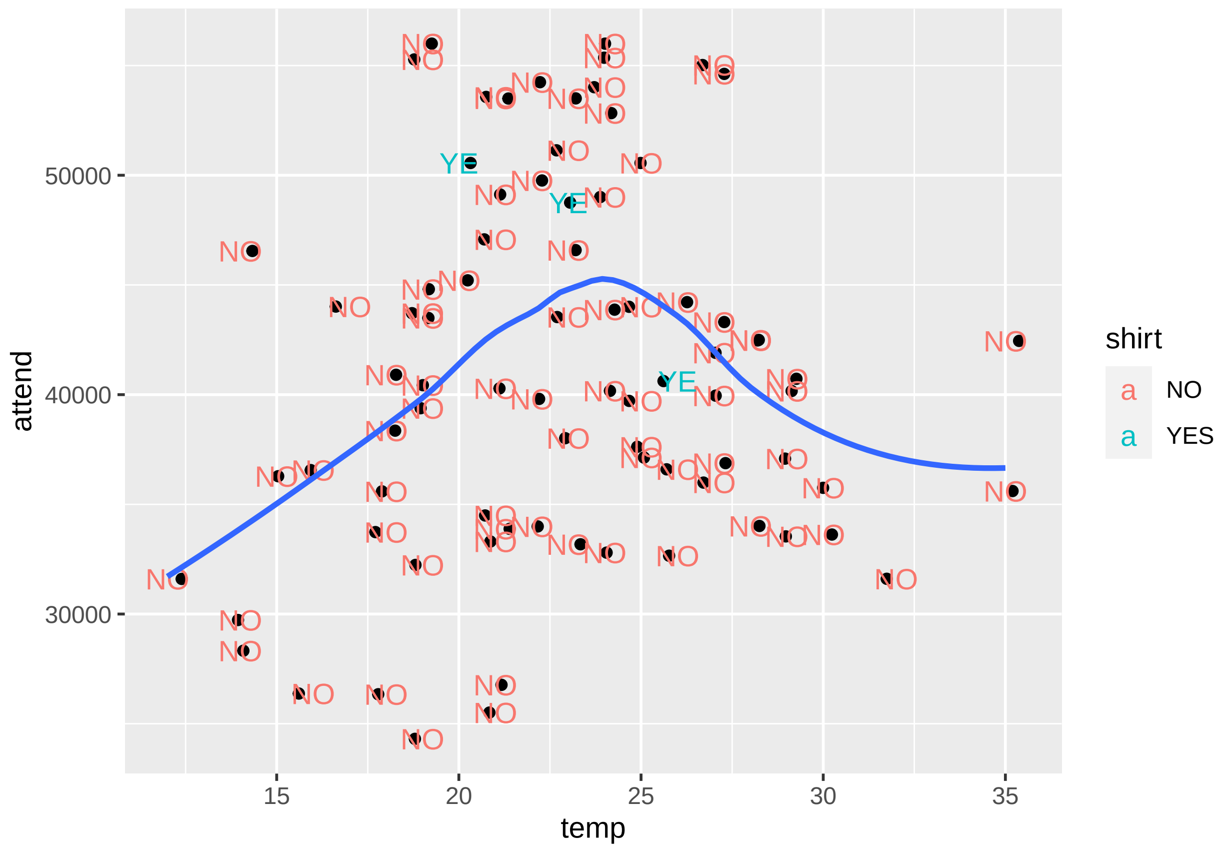
shirt & temperature

It is possible shirt promotion has different effects on different temperatures. Let's examine this.

```
ggplot(data = events, aes(temp, attend)) +
  geom_jitter() +
  geom_text(data = subset(events, shirt %in% c("YES", "NO")),
            aes(label = str_sub(shirt, 1, 2), col = shirt)) +
```

```
  geom_smooth(se = FALSE)
```

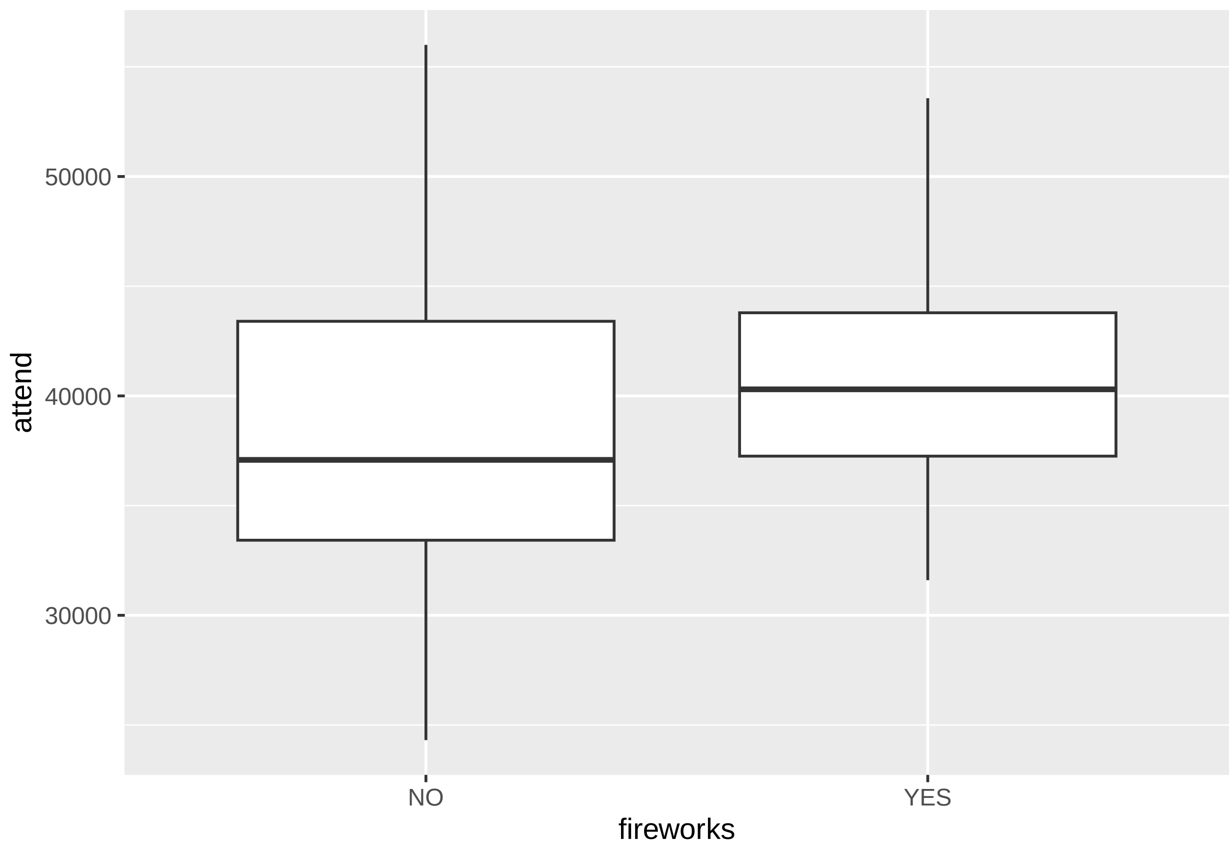
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The plot shows that when temperature is low, the shirt promotion results in more attendance. However, we have only three matches with shirt promotion, so it is not reasonable to completely trust on assumptions made from this plot.

fireworks

```
ggplot(data=events[bobblehead=="NO" & cap=="NO" & shirt=="NO"], aes(fireworks, attend)) +
  geom_boxplot()
```



It seems fireworks does not have an effect on attend as nearly all observations of group YES matches with group NO.

```
t.test(events[bobblehead=="NO" & cap=="NO" & shirt=="NO" & fireworks=="YES", attend],
       events[bobblehead=="NO" & cap=="NO" & shirt=="NO" & fireworks=="NO", attend])
```

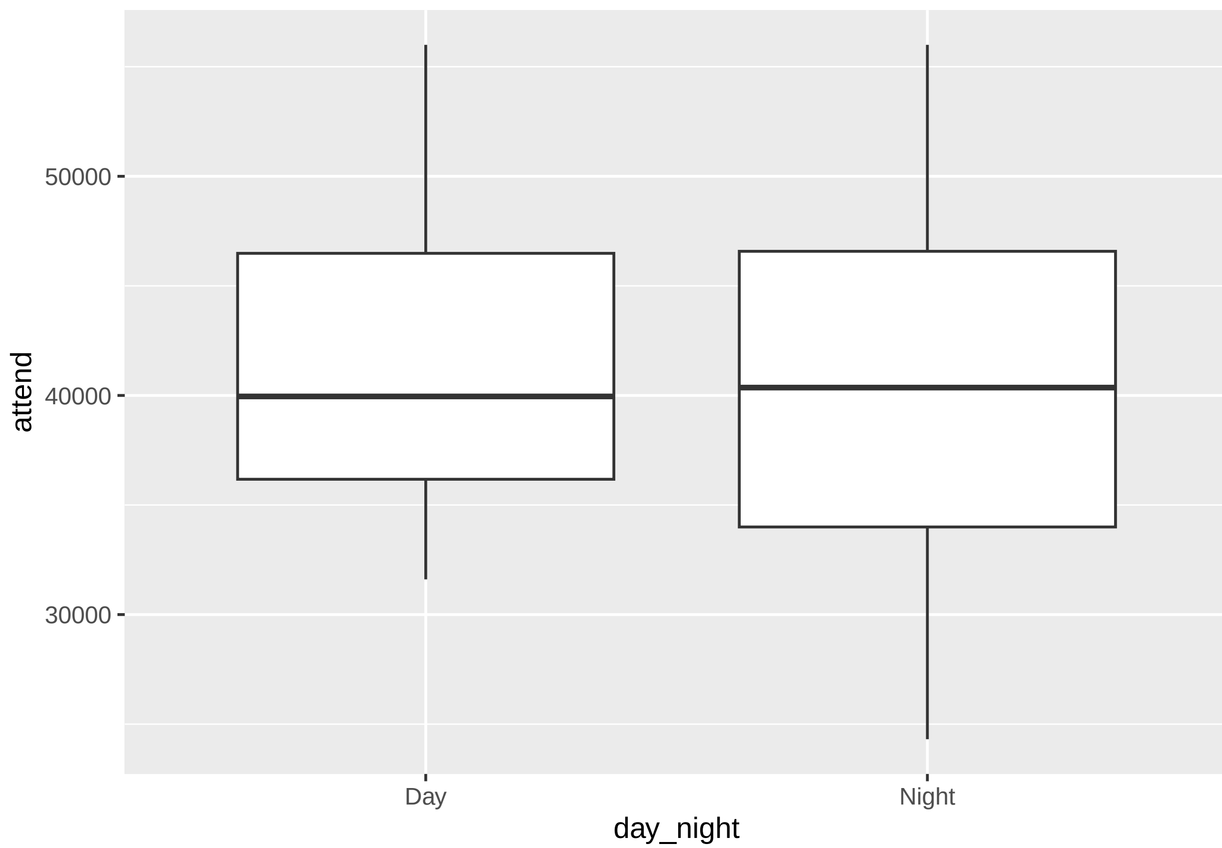
```
##
##  Welch Two Sample t-test
##
## data:  events[bobblehead == "NO" & cap == "NO" & shirt == "NO" & fireworks == "YES", attend] and eve
## t = 1.5463, df = 26.155, p-value = 0.1341
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -946.3174 6699.8748
## sample estimates:
## mean of x mean of y
##  41077.86  38201.08
```

The t-test gives the p-value of 0.1341. Hence, we do not see a statistically significant difference between the average attendance of the games played under fireworks or not.

day_night

We will check if there is an association between attendance and whether the game is played in day light or night.

```
ggplot(data=events, aes(day_night, attend)) +
geom_boxplot()
```



The boxplot does not suggest a strong difference.

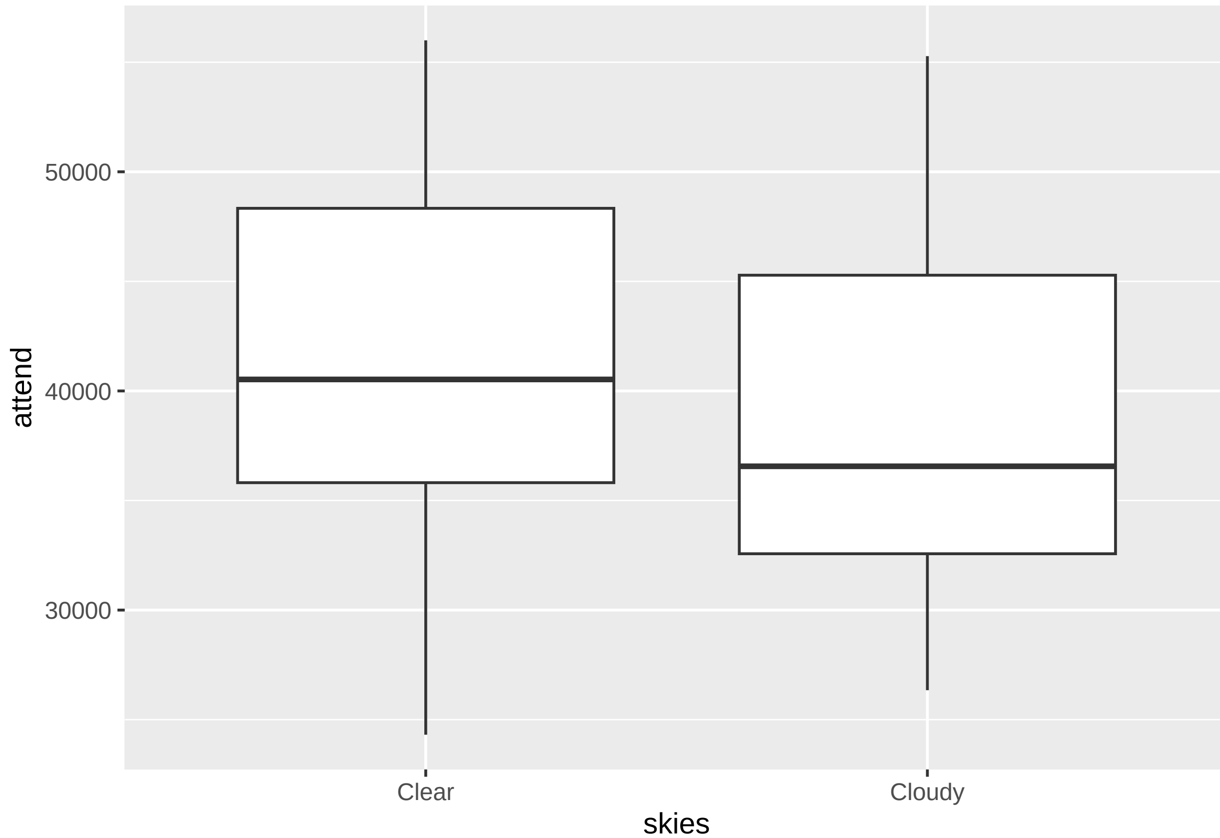
```
t.test(x=events[day_night=="Day", attend],
       y=events[day_night=="Night", attend])
```

```
##
##  Welch Two Sample t-test
##
## data:  events[day_night == "Day", attend] and events[day_night == "Night", attend]
## t = 0.42851, df = 23.62, p-value = 0.6722
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3531.652  5380.397
## sample estimates:
## mean of x mean of y
##  41793.27  40868.89
```

Since p-value (0.67) is large (greater than 0.05), we cannot reject null, which means there is no statistical difference between average attendance of games played in day and night.

skies

```
  ggplot(data=events, aes(skies, attend)) +
  geom_boxplot()
```



The plot does not show an important difference.

```
t.test(events[skies=="Clear", attend],
       events[skies=="Cloudy", attend])
```

```
##
##  Welch Two Sample t-test
##
## data:  events[skies == "Clear", attend] and events[skies == "Cloudy", attend]
## t = 1.2868, df = 27.664, p-value = 0.2088
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1741.315  7617.103
## sample estimates:
## mean of x mean of y
##  41729.21  38791.32
```

The t-test backs up our hypothesis. It says there is no statistically significant difference between the average attendance of the games played under clear and cloudy skies.

skies & day_night

It is reasonable to suggest that skies and day_night variables are related because a day with a clear sky probably has a different effect on attendance than a cloudy day.

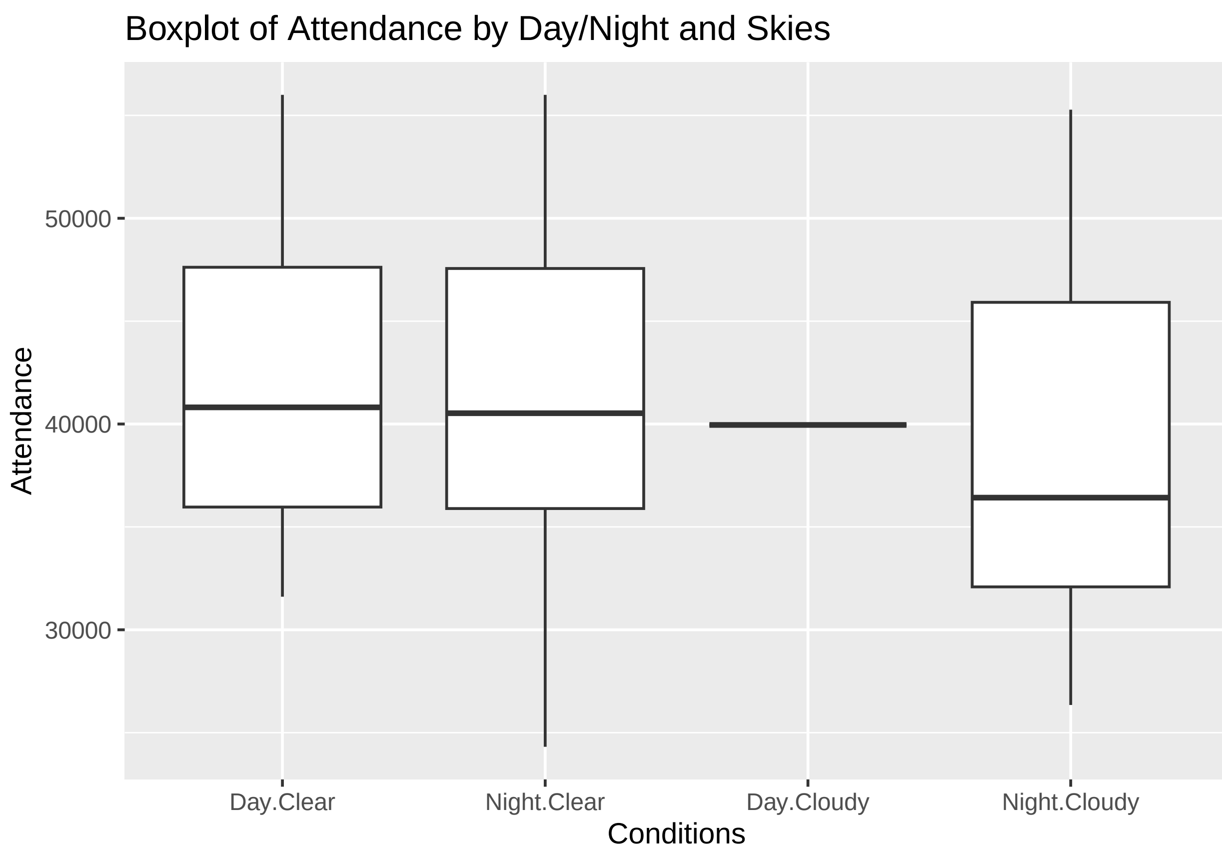
```
unique(events[, .(day_night, skies)])
```

```
##    day_night  skies
```

```
## 1:       Day  Clear
## 2:     Night Cloudy
## 3:     Night  Clear
## 4:       Day Cloudy
```

So, those combinations' effects on attendance is going to be analyzed.

```r
ggplot(events, aes(x = interaction(day_night, skies), y = attend)) +
  geom_boxplot() +
  labs(x = "Conditions", y = "Attendance") +
  ggtitle("Boxplot of Attendance by Day/Night and Skies")
```



The boxplot does not show a significant difference in attendances under different conditions. We can apply an ANOVA test to back-up or falsify our hypothesis.

```r
# perform ANOVA test
model <- aov(attend ~ day_night * skies, data = events)

# summarize ANOVA results
summary(model)
```

```
##                 Df     Sum Sq   Mean Sq F value Pr(>F)
## day_night        1   10443460  10443460   0.149  0.700
## skies            1  116371868 116371868   1.665  0.201
## day_night:skies  1     829887    829887   0.012  0.914
## Residuals       77 5380287672  69873866
```

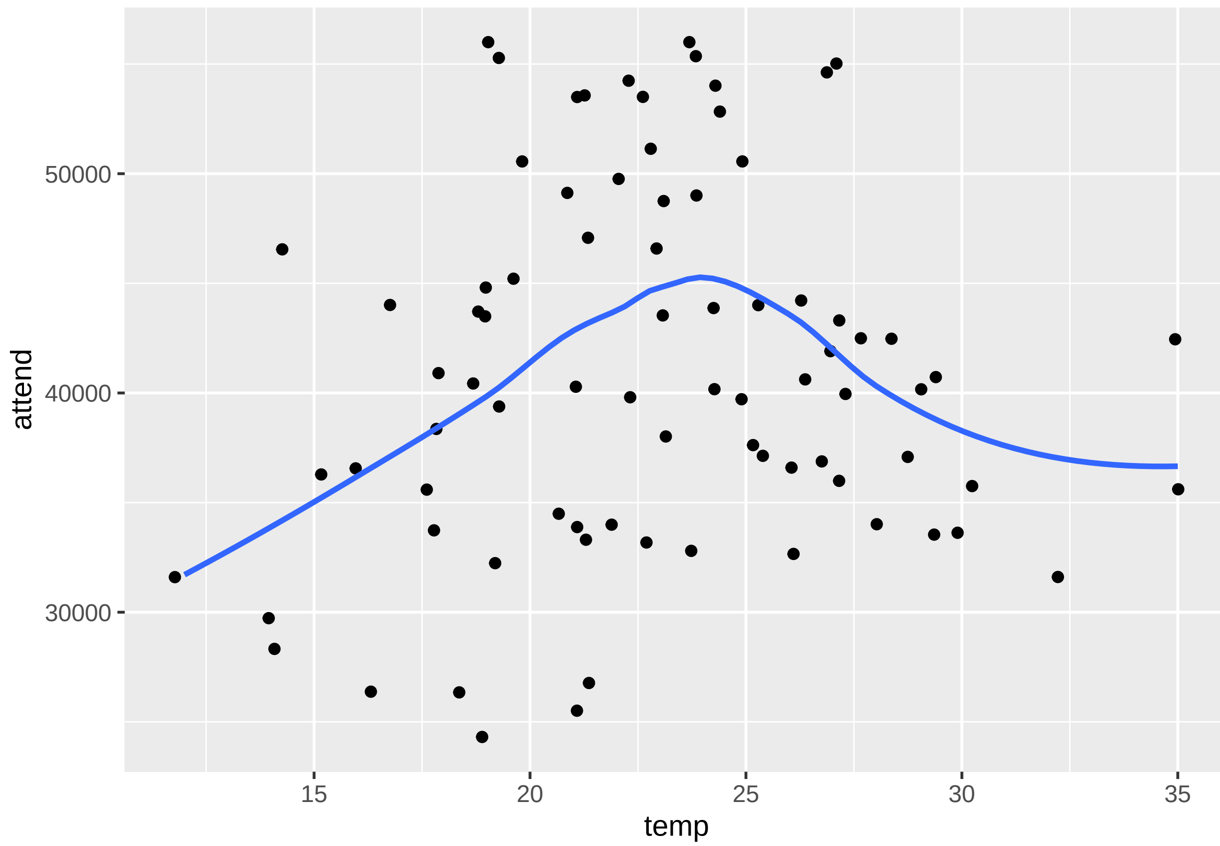
Based on this output, we can see that none of the terms in the model are statistically significant at the significance level of 0.05. This means that there is no evidence of a significant difference in attendance based on the day_night and skies conditions or their interaction.

temperature

Now, we will check if there is an association between attendance and temperature.

```
ggplot(data= events, aes(temp, attend)) +
geom_jitter() +
geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



From the loess fit, it seems that attendance is positively correlated with temperature until 23 celcius. After that point, they seems to be negatively correlated.

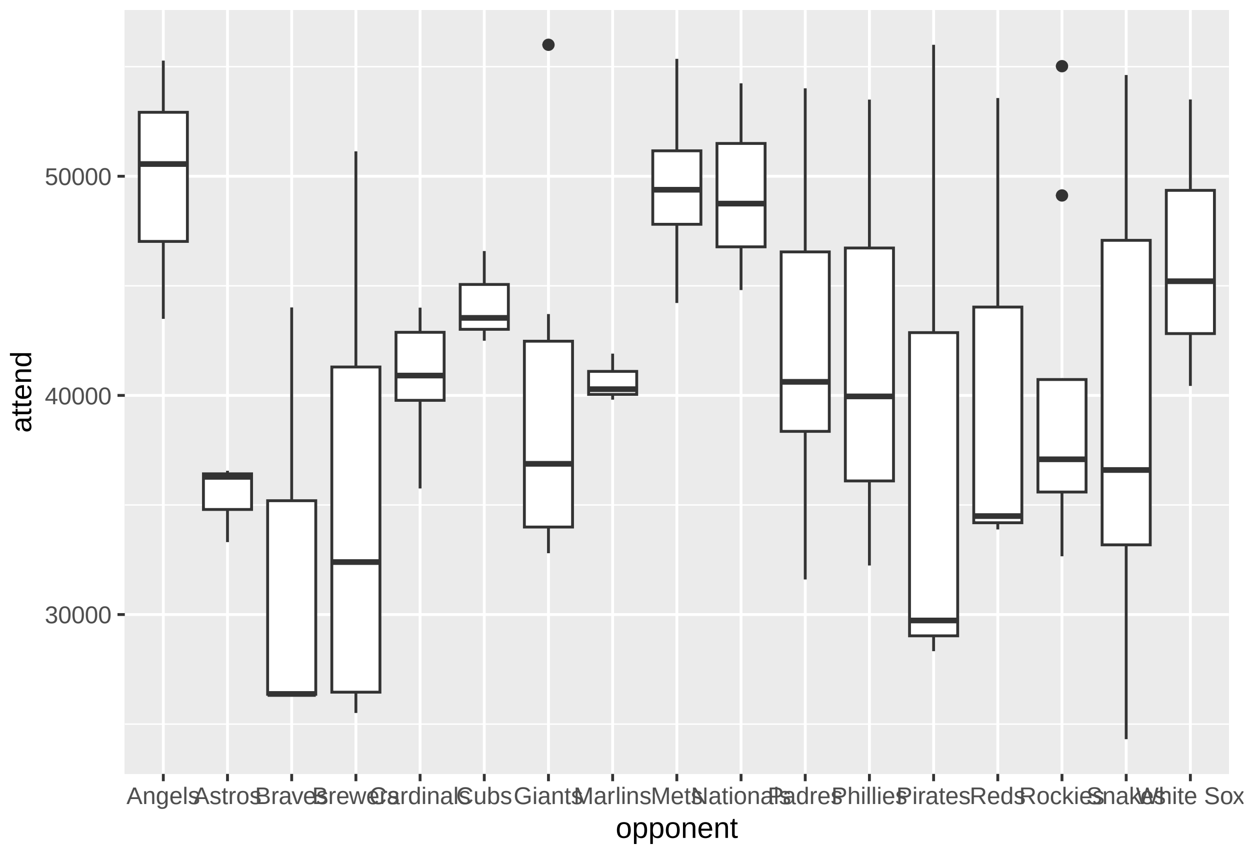
opponent

```
events[, .(number_of_games= .N,
           mean_attend= mean(attend)),
       opponent][order(-number_of_games)]
```

```
##      opponent number_of_games mean_attend
##   1:   Padres               9    42092.22
##   2:   Giants               9    39296.33
##   3:  Rockies               9    39631.22
##   4:   Snakes               9    39315.44
##   5: Cardinals              7    40853.29
```

```
## 6:   Brewers          4    35358.75
## 7:      Mets          4    49586.25
## 8:   Pirates          3    38019.00
## 9:    Braves          3    32245.00
## 10: Nationals         3    49267.33
## 11:    Astros         3    35383.33
## 12:    Angels         3    49777.33
## 13: White Sox         3    46382.00
## 14:      Reds         3    40649.00
## 15:  Phillies         3    41897.00
## 16:      Cubs         3    44206.67
## 17:   Marlins         3    40665.33
```

```
ggplot(data=events, aes(opponent, attend)) +
geom_boxplot()
```



To see whether there is significant difference in the mean attendance values of the groups due to opponent, ANOVA test will be applied.

```
anova_model <- aov(attend ~ opponent, data = events)
summary(anova_model)
```

```
##             Df     Sum Sq  Mean Sq F value Pr(>F)
## opponent    16 1409757018 88109814   1.376  0.183
## Residuals   64 4098175870 64033998
```

The p-value is 0.183 which is greater than 0.05 . It means there is not any opponent whose attend values are significantly different from the others.

month

```r
events[, .(numberOfMatch=.N), month][order(-numberOfMatch)]
```

```
##    month numberOfMatch
## 1:   MAY            18
## 2:   AUG            15
## 3:   APR            12
## 4:   JUL            12
## 5:   SEP            12
## 6:   JUN             9
## 7:   OCT             3
```

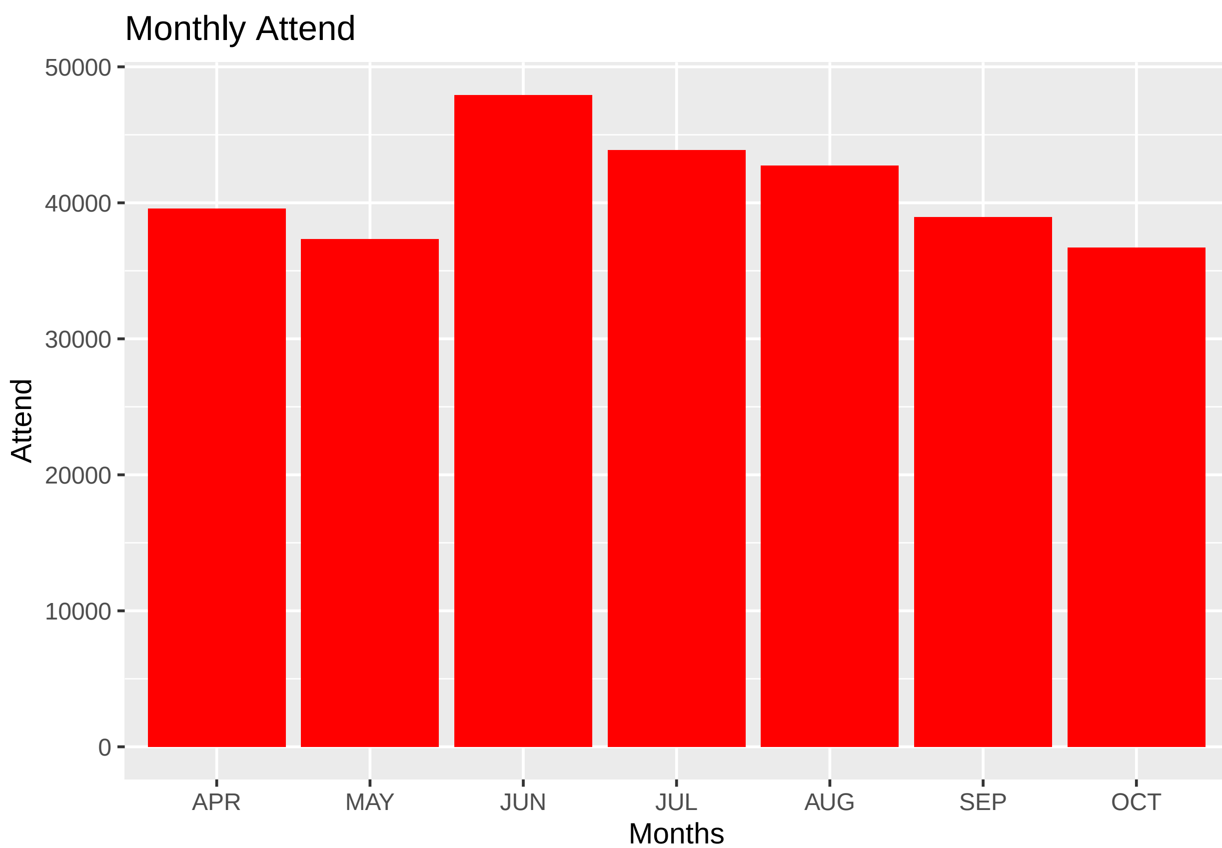
There are quite less matches in october compared to other months. Let's check mean attend in each month.

```r
events[, .(meanAttend=mean(attend)), month][order(-meanAttend)]
```

```
##    month meanAttend
## 1:   JUN   47940.44
## 2:   JUL   43884.25
## 3:   AUG   42751.53
## 4:   APR   39591.92
## 5:   SEP   38955.08
## 6:   MAY   37345.72
## 7:   OCT   36703.67
```

It seems some months June and July have bigger attendance values compared to others.

```r
ggplot(data=events[, .(meanAttend=mean(attend)), month],
       aes(month, meanAttend)) +
  geom_bar(stat="identity", fill="red") +
  labs(title="Monthly Attend", x="Months", y="Attend")
```




```
anova_model <- aov(attend ~ month, data = events)
summary(anova_model)
```

```
##             Df     Sum Sq   Mean Sq F value Pr(>F)
## month        6  948958117 158159686   2.567 0.0258 *
## Residuals   74 4558974770  61607767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA suggests that least one of the month's mean is significantly different from the others. To determine which months are significantly different from the rest, we'll apply **Tukey's HSD (honestly significant difference) test**. This test compares all pairwise differences between the month means and adjusts for multiple comparisons to control the family-wise error rate.

```
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = attend ~ month, data = events)
##
## $month
##               diff         lwr       upr     p adj
## MAY-APR -2246.1944 -11112.1779  6619.789 0.9872921
## JUN-APR  8348.5278  -2141.8454 18838.901 0.2083242
## JUL-APR  4292.3333  -5419.8650 14004.532 0.8310207
## AUG-APR  3159.6167  -6054.1837 12373.417 0.9430190
## SEP-APR  -636.8333 -10349.0316  9075.365 0.9999945
```

```
## OCT-APR  -2888.2500 -18244.5839 12468.084 0.9974623
## JUN-MAY  10594.7222    882.5239 20306.921 0.0235396
## JUL-MAY   6538.5278  -2327.4557 15404.511 0.2898249
## AUG-MAY   5405.8111  -2911.2186 13722.841 0.4421814
## SEP-MAY   1609.3611  -7256.6224 10475.345 0.9979181
## OCT-MAY   -642.0556 -15477.6835 14193.572 0.9999995
## JUL-JUN  -4056.1944 -14546.5676  6434.179 0.9024009
## AUG-JUN  -5188.9111 -15219.6264  4841.804 0.7028001
## SEP-JUN  -8985.3611 -19475.7343  1505.012 0.1421597
## OCT-JUN -11236.7778 -27096.7312  4623.176 0.3366731
## AUG-JUL  -1132.7167 -10346.5170  8081.084 0.9997754
## SEP-JUL  -4929.1667 -14641.3650  4783.032 0.7210021
## OCT-JUL  -7180.5833 -22536.9172  8175.751 0.7908461
## SEP-AUG  -3796.4500 -13010.2503  5417.350 0.8723664
## OCT-AUG  -6047.8667 -21093.9396  8998.206 0.8848659
## OCT-SEP  -2251.4167 -17607.7505 13104.917 0.9993792
```

From this output, we can see that the mean attendance for month June is significantly different from month May ($p < 0.05$), but there is no significant difference between other groups ($p > 0.05$).

day

Maybe the majority of the citizens have a payment day within first week of a month. This may result in an increased attendance in the first week of a month.

```
events[, .(mean_attendance = mean(attend)), day][order(-mean_attendance)]
```

```
##      day mean_attendance
##  1:   10        56000.00
##  2:   21        56000.00
##  3:   17        53501.00
##  4:    7        49368.50
##  5:   29        47594.25
##  6:    4        46925.67
##  7:    5        46527.50
##  8:   28        44599.25
##  9:   13        42280.20
## 10:   24        41909.50
## 11:   15        41606.40
## 12:   14        41260.50
## 13:   20        40441.50
## 14:   18        40430.50
## 15:    1        40392.75
## 16:   22        40173.00
## 17:   19        39383.00
## 18:   26        39234.00
## 19:   12        39114.00
## 20:   31        39075.67
## 21:   27        39056.50
## 22:   30        38626.80
## 23:   11        38626.33
## 24:   16        37734.00
## 25:    3        36243.75
## 26:    2        36191.00
## 27:    8        34941.50
```

```
## 28:  25        34304.00
## 29:   9        33993.00
## 30:   6        32659.00
## 31:  23        26376.00
##     day mean_attendance
```

It seems highly random.

```r
events[, day := factor(day)]
anova_model <- aov(attend ~ day, data = events)
summary(anova_model)
```

```
##             Df     Sum Sq  Mean Sq F value Pr(>F)
## day         30 2125645147 70854838   1.047  0.433
## Residuals   50 3382287740 67645755
```

ANOVA suggests that none one of the month's mean is significantly different from the others.

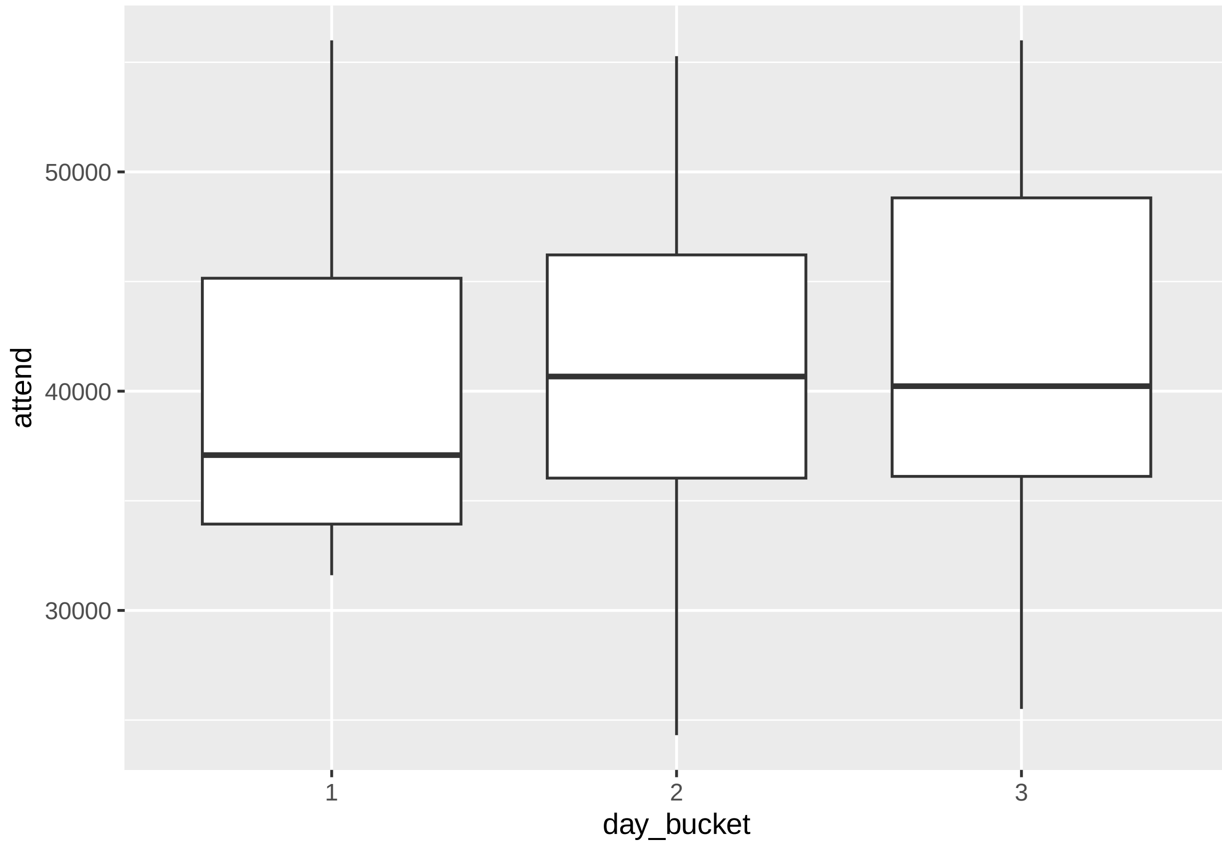
Now we are going to divide day into 3 buckets, meaning 0-10, 10-20 & 20+. Then, we will examine attendance patterns of those buckets.

```r
events[, day := as.integer(day)]
events[, day_bucket := 3]
events[day <= 20, day_bucket := 2]
events[day <= 10, day_bucket := 1]
head(events[, .(day, day_bucket)])
```

```
##    day day_bucket
## 1:  10          1
## 2:  11          2
## 3:  12          2
## 4:  13          2
## 5:  14          2
## 6:  15          2
```

Now let's check if there are serious differences in attendance for different day buckets.

```r
events[, day_bucket := factor(day_bucket)]

ggplot(data=events, aes(day_bucket, attend)) +
geom_boxplot()
```



The plot shows that attendance does not change with days. To statistically support our idea, we can apply ANOVA test.

```
anova_model <- aov(attend ~ day_bucket, data = events)
summary(anova_model)
```

```
##             Df     Sum Sq  Mean Sq F value Pr(>F)
## day_bucket   2    3243723  1621861   0.023  0.977
## Residuals   78 5504689165 70572938
```

ANOVA suggests that there is not any bucket that has significantly different attendance from others.

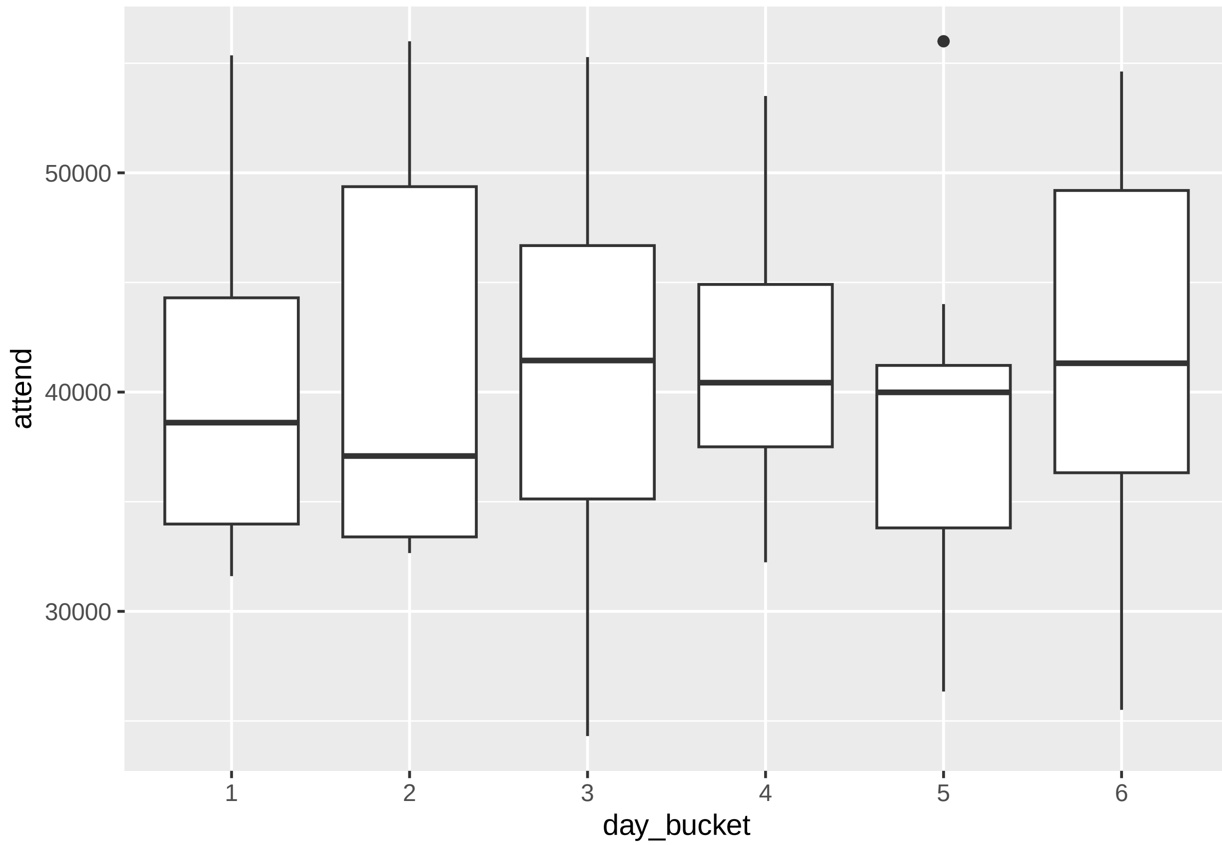
Maybe the issue lies in the width of the buckets. We can analyze the data in 5 days intervals instead of 10.

```
events[, day_bucket := NULL]
events[, day_bucket := 6]
events[day <= 25, day_bucket := 5]
events[day <= 20, day_bucket := 4]
events[day <= 15, day_bucket := 3]
events[day <= 10, day_bucket := 2]
events[day <= 5, day_bucket := 1]

head(events[, .(day, day_bucket)])
```

```
##    day day_bucket
## 1:  10          2
## 2:  11          3
## 3:  12          3
## 4:  13          3
```

```
## 5:  14          3
## 6:  15          3
```

```r
events[, day_bucket := factor(day_bucket)]

ggplot(data=events, aes(day_bucket, attend)) +
geom_boxplot()
```



Still the plot shows that there is not any significance difference among buckets.

```r
anova_model <- aov(attend ~ day_bucket, data = events)
summary(anova_model)
```

```
##             Df     Sum Sq  Mean Sq F value Pr(>F)
## day_bucket   5   74841595 14968319   0.207  0.959
## Residuals   75 5433091293 72441217
```

ANOVA test supports our inference. Of course there may be other relationships, for example between bobblehead and days and those ones will be examined in the model development part.

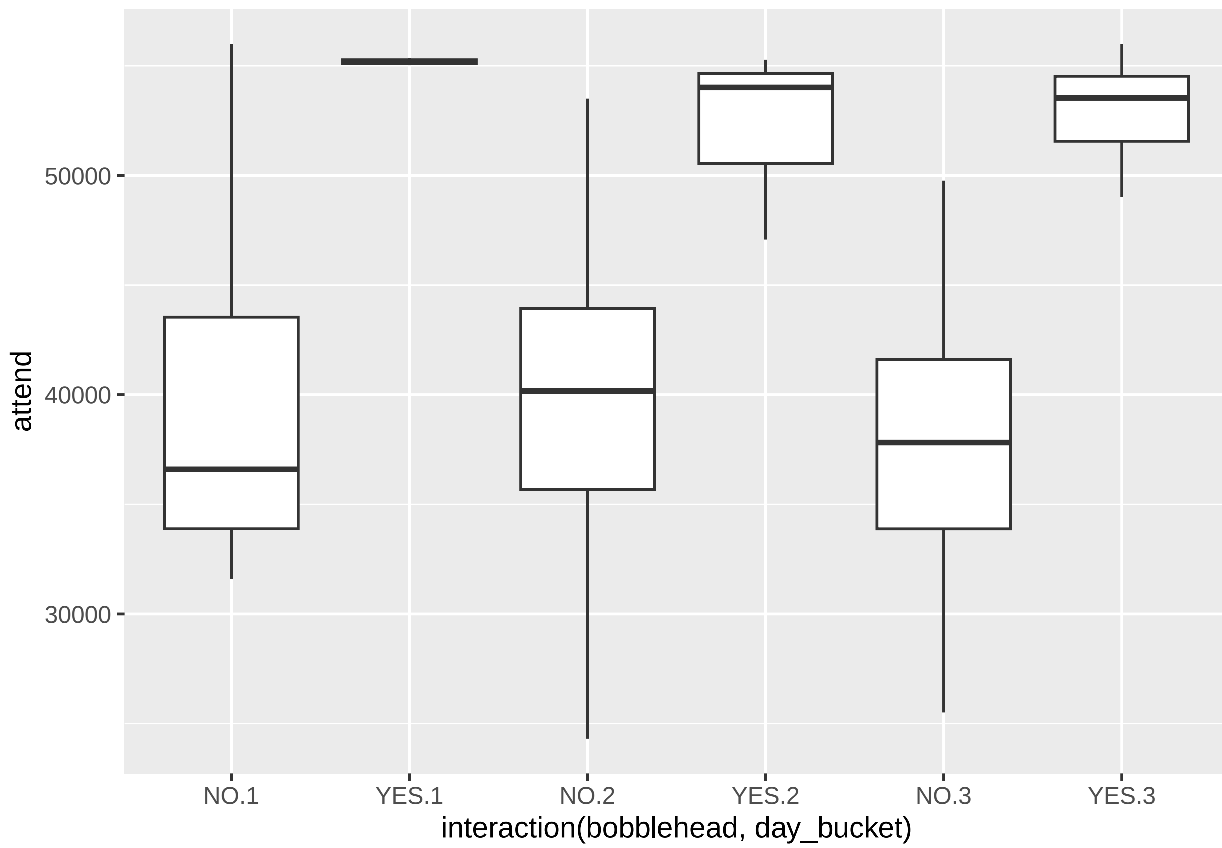
bobblehead & day

It is possible that bobblehead promotions and days of a month have an association. Let's check that.

```r
events[, day := as.integer(day)]
events[, day_bucket := 3]
events[day <= 20, day_bucket := 2]
events[day <= 10, day_bucket := 1]
```

Now let's check if there are serious differences in attendance for different day buckets.

```
ggplot(events, aes(x = interaction(bobblehead, day_bucket), y = attend)) +
  geom_boxplot()
```



There seems to be no relation between days of a month and bobblehead.

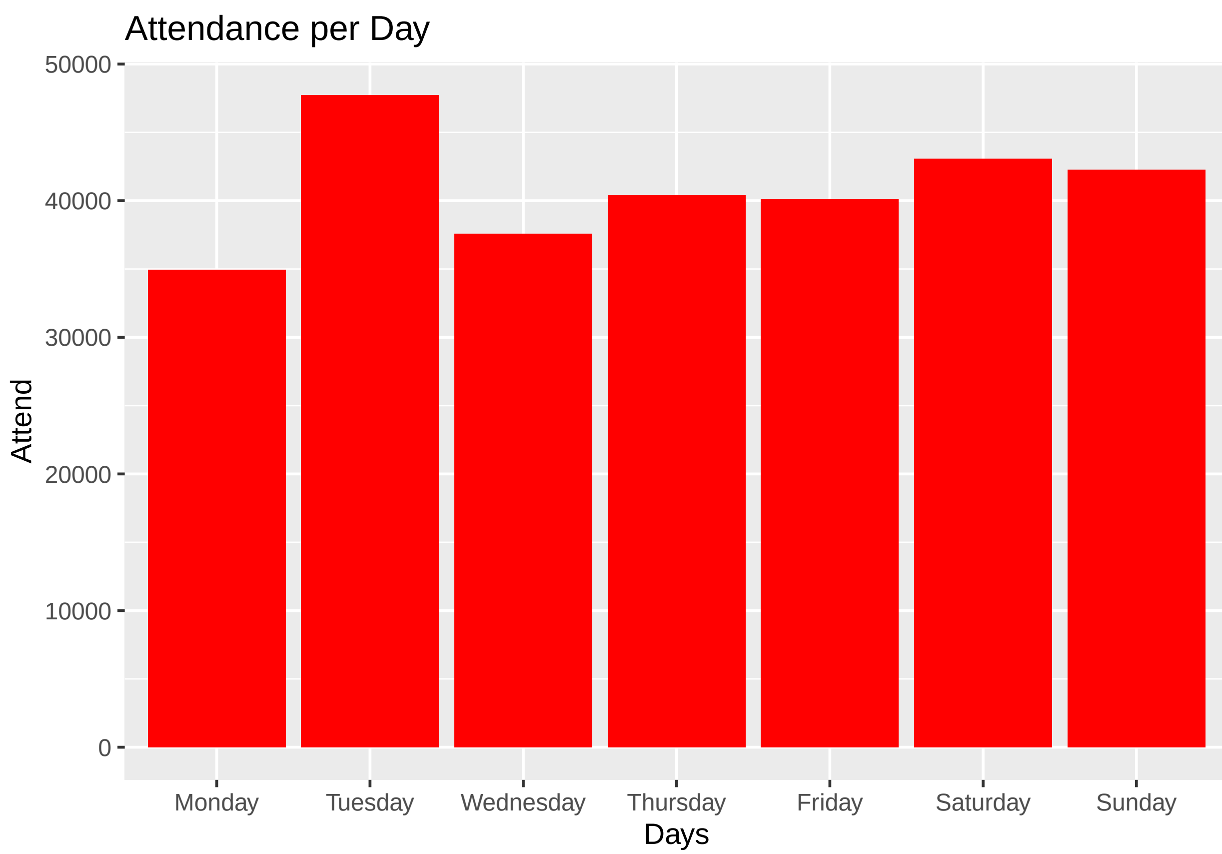
day_of_week

```
events[, .(numberOfMatch=.N), day_of_week][order(-numberOfMatch)]
```

```
##    day_of_week numberOfMatch
## 1:     Tuesday            13
## 2:      Friday            13
## 3:    Saturday            13
## 4:      Sunday            13
## 5:   Wednesday            12
## 6:      Monday            12
## 7:    Thursday             5
```

There are quite less matches in thursday compared to other days. Let's check mean attendance in each day.

```
ggplot(data=events[, .(meanAttend=mean(attend)), day_of_week],
       aes(day_of_week, meanAttend)) +
  geom_bar(stat="identity", fill="red") +
  labs(title="Attendance per Day", x="Days", y="Attend")
```



Only Monday and Tuesday has a huge difference as seen in the above bar plot.

```
anova_model <- aov(attend ~ day_of_week, data = events)
summary(anova_model)
```

```
##             Df     Sum Sq   Mean Sq F value  Pr(>F)
## day_of_week  6 1256219950 209369992   3.644 0.00319 **
## Residuals   74 4251712937  57455580
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA suggests that least one of the day's mean is significantly different from the others. To determine which days are significantly different from the rest, we'll apply Tukey's HSD (honestly significant difference) test. This test compares all pairwise differences between the day means and adjusts for multiple comparisons to control the family-wise error rate.

```
TukeyHSD(anova_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = attend ~ day_of_week, data = events)
##
## $day_of_week
##                           diff        lwr        upr     p adj
## Tuesday-Monday      12775.5641   3578.499 21972.6289 0.0013332
## Wednesday-Monday     2619.5000  -6759.703 11998.7025 0.9790113
## Thursday-Monday      5441.7333  -6787.251 17670.7173 0.8264855
```

```
## Friday-Monday        5151.2564  -4045.808 14348.3212 0.6197791
## Saturday-Monday      8107.2564  -1089.808 17304.3212 0.1201717
## Sunday-Monday        7303.1795  -1893.885 16500.2443 0.2105118
## Wednesday-Tuesday  -10156.0641 -19353.129  -958.9993 0.0209502
## Thursday-Tuesday    -7333.8308 -19423.686  4756.0248 0.5269152
## Friday-Tuesday      -7624.3077 -16635.554  1386.9386 0.1522316
## Saturday-Tuesday    -4668.3077 -13679.554  4342.9386 0.7013870
## Sunday-Tuesday      -5472.3846 -14483.631  3538.8617 0.5255747
## Thursday-Wednesday   2822.2333  -9406.751 15051.2173 0.9922411
## Friday-Wednesday     2531.7564  -6665.308 11728.8212 0.9804977
## Saturday-Wednesday   5487.7564  -3709.308 14684.8212 0.5467123
## Sunday-Wednesday     4683.6795  -4513.385 13880.7443 0.7177999
## Friday-Thursday      -290.4769 -12380.333 11799.3787 1.0000000
## Saturday-Thursday    2665.5231  -9424.333 14755.3787 0.9939311
## Sunday-Thursday      1861.4462 -10228.409 13951.3018 0.9991787
## Saturday-Friday      2956.0000  -6055.246 11967.2463 0.9537334
## Sunday-Friday        2151.9231  -6859.323 11163.1694 0.9906918
## Sunday-Saturday      -804.0769  -9815.323  8207.1694 0.9999656
```

From this output, we can see that the mean attendance for Tuesday-Monday and Wednesday-Tuesday are significantly different from each other ($p < 0.05$), but there is no significant difference between other pairs ($p > 0.05$).

day_of_week & bobblehead

It is stated that attendance in Tuesday is significantly higher than Monday and Tuesday, but maybe it is because another variable. Now, we will check interacted variables' relationship.

```
# Create a contingency table of day_of_week and bobblehead promotion
cont_table <- table(events$day_of_week, events$bobblehead)
cont_table
```

```
##
##             NO YES
##   Monday    12   0
##   Tuesday    7   6
##   Wednesday 12   0
##   Thursday   3   2
##   Friday    13   0
##   Saturday  11   2
##   Sunday    12   1
```

We resulted that bobblehead promotion has a statistically significant effect on the attendance and **contingency table** shows that days with high attendance have bobblehead promotions and days with less attendance have not bobblehead promotions. So, days and bobblehead promotions seem to be associated. We can check this by **chi-squared test of independence**.

```
chisq.test(cont_table)
```

```
## Warning in chisq.test(cont_table): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 20.961, df = 6, p-value = 0.001864
```

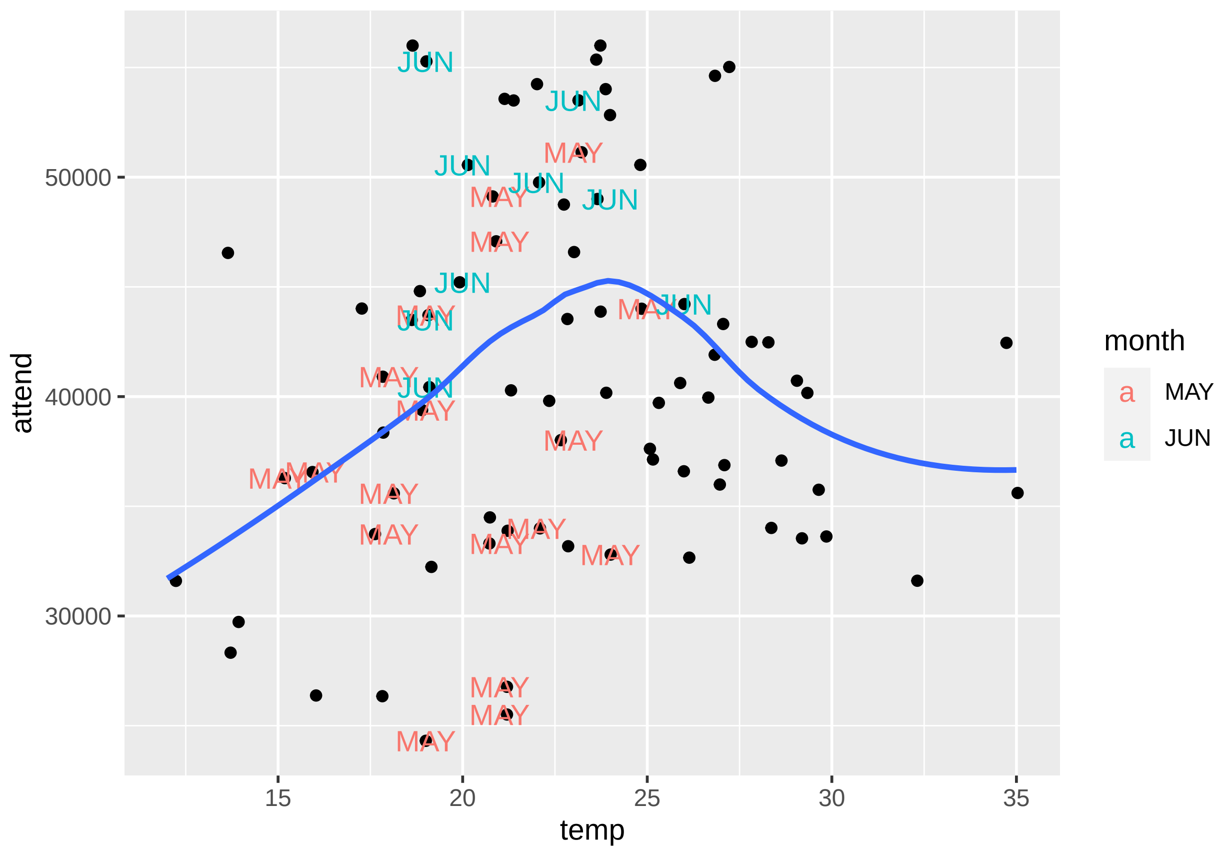
p-value of 0.001864 indicates that the observed association between the variables (days of the week and bobblehead promotions) is statistically significant.

month & temperature

We concluded that the mean attendance for month June is significantly different from month May ($p < 0.05$), but there is no significant difference between other groups ($p > 0.05$). Maybe, this difference is due to temperature or another variable. Let's check this.

```
ggplot(data = events, aes(temp, attend)) +
  geom_jitter() +
  geom_text(data = subset(events, month %in% c("MAY", "JUN")),
            aes(label = str_sub(month, 1, 3), col = month)) +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



First of all, when the relationship between temperature and attendance is examined, people tend to go to stadium when the weather is mild, namely not cold and not hot. So, the most attendance happen within the range of IQR.

```
IQR_upper <- quantile(events[, temp], .75)
IQR_lower <- quantile(events[, temp], .25)
paste0("Q1 - Q3 = ", IQR_lower, " - ", IQR_upper)
```

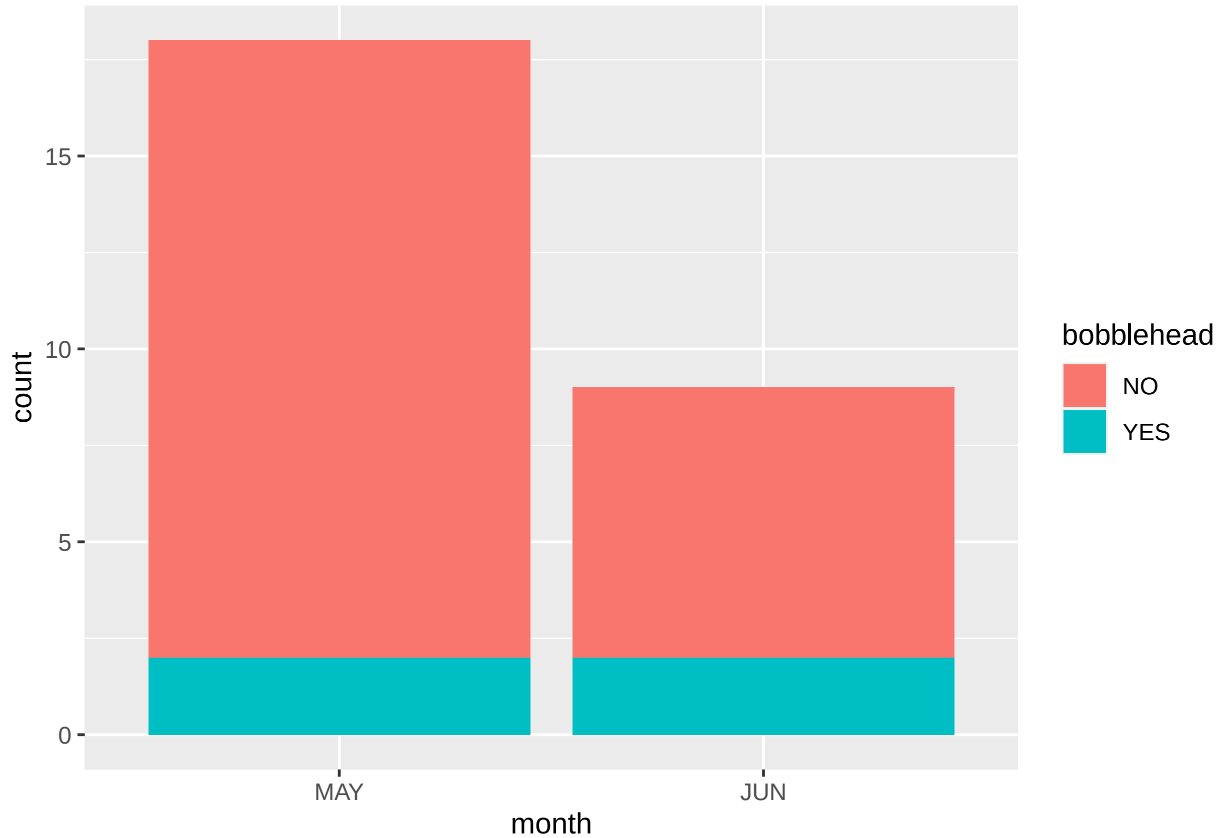
```
## [1] "Q1 - Q3 = 19 - 26"
```

However, in may, although the weather is in this range, people are not tend to go to the stadium. So, obviously there are other factors that influence attendance. Let's observe if there is a bobblehead promotion on these matches.

```
events[month %in% c("MAY", "JUN"), .(month, temp, bobblehead, attend)]
```

```
##     month temp bobblehead attend
##  1:   MAY   19         NO  43713
##  2:   MAY   24         NO  32799
##  3:   MAY   22         NO  33993
##  4:   MAY   18         NO  35591
##  5:   MAY   18         NO  33735
##  6:   MAY   21         NO  49124
##  7:   MAY   19         NO  24312
##  8:   MAY   21        YES  47077
##  9:   MAY   18         NO  40906
## 10:   MAY   19         NO  39383
## 11:   MAY   25         NO  44005
## 12:   MAY   15         NO  36283
## 13:   MAY   16         NO  36561
## 14:   MAY   21         NO  33306
## 15:   MAY   23         NO  38016
## 16:   MAY   23        YES  51137
## 17:   MAY   21         NO  25509
## 18:   MAY   21         NO  26773
## 19:   JUN   20         NO  50559
## 20:   JUN   19        YES  55279
## 21:   JUN   19         NO  43494
## 22:   JUN   19         NO  40432
## 23:   JUN   20         NO  45210
## 24:   JUN   23         NO  53504
## 25:   JUN   24        YES  49006
## 26:   JUN   22         NO  49763
## 27:   JUN   26         NO  44217
##     month temp bobblehead attend
```

```
ggplot(data=events[month %in% c("MAY", "JUN")], aes(month)) +
geom_bar(aes(fill=bobblehead))
```



Almost all days of JUNE have nice weather and also high attendances although not all of them have bobblehead promotion. However, when same conditions are applied in MAY, the attendance is quite low. So, it is not only the bobblehead, month and weather that affect attendance. There may be other variables or interaction terms.

step function

We can use step function by inputting some relevant variables.

```
# Fit the full model with all possible predictors, including interaction terms
events[, day := factor(day)]

model_full <- lm(attend ~ (month+
                             day+
                             day_of_week+
                             opponent+
                             temp+
                             pmax(0, temp - 23)+
                             day_night+
                             cap+
                             shirt+
                             bobblehead+
                             bobblehead*day_of_week+
                             skies*fireworks+
                             skies*day_night*fireworks+
                             skies*day_night*bobblehead+
                             cap*shirt*temp+
```

```r
                              opponent*fireworks+
                              fireworks*day_night*skies), data = events)

# Use stepwise selection to find the best model with highest adjusted R-squared, including interaction
model_best <- step(model_full, direction = "backward", k = log(nrow(events)))
```

```
## Start:  AIC=946.66
## attend ~ (month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + bobblehead *
##     day_of_week + skies * fireworks + skies * day_night * fireworks +
##     skies * day_night * bobblehead + cap * shirt * temp + opponent *
##     fireworks + fireworks * day_night * skies)
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + skies:fireworks + day_night:skies +
##     day_night:fireworks + bobblehead:skies + day_night:bobblehead +
##     cap:shirt + temp:cap + temp:shirt + opponent:fireworks +
##     day_night:skies:fireworks + day_night:bobblehead:skies
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + skies:fireworks + day_night:skies +
##     day_night:fireworks + bobblehead:skies + day_night:bobblehead +
##     cap:shirt + temp:cap + temp:shirt + opponent:fireworks +
##     day_night:skies:fireworks
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + skies:fireworks + day_night:skies +
##     day_night:fireworks + bobblehead:skies + day_night:bobblehead +
##     cap:shirt + temp:cap + temp:shirt + opponent:fireworks
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + skies:fireworks + day_night:skies +
##     day_night:fireworks + bobblehead:skies + day_night:bobblehead +
##     temp:cap + temp:shirt + opponent:fireworks
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + skies:fireworks + day_night:skies +
##     day_night:fireworks + bobblehead:skies + temp:cap + temp:shirt +
##     opponent:fireworks
```

```
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + skies:fireworks + day_night:skies +
##     bobblehead:skies + temp:cap + temp:shirt + opponent:fireworks
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + skies:fireworks + bobblehead:skies +
##     temp:cap + temp:shirt + opponent:fireworks
##
##
## Step:  AIC=946.66
## attend ~ month + day + day_of_week + opponent + temp + pmax(0,
##     temp - 23) + day_night + cap + shirt + bobblehead + skies +
##     fireworks + day_of_week:bobblehead + bobblehead:skies + temp:cap +
##     temp:shirt + opponent:fireworks
##
##                         Df  Sum of Sq        RSS     AIC
## <none>                                    125671  946.66
## - temp:shirt             1     829117     954788 1106.52
## - pmax(0, temp - 23)     1   17420085   17545756 1342.32
## - temp:cap               1   91502963   91628634 1476.20
## - day_night              1  118309709  118435380 1496.99
## - day_of_week:bobblehead  2  136867993  136993664 1504.39
## - bobblehead:skies       1  133868730  133994401 1506.99
## - month                  5  231934589  232060260 1533.90
## - opponent:fireworks     8  313746482  313872153 1545.17
## - day                   29 1030552977 1030678648 1549.20
```

```
summary(model_best)
```

```
##
## Call:
## lm(formula = attend ~ month + day + day_of_week + opponent +
##     temp + pmax(0, temp - 23) + day_night + cap + shirt + bobblehead +
##     skies + fireworks + day_of_week:bobblehead + bobblehead:skies +
##     temp:cap + temp:shirt + opponent:fireworks, data = events)
##
## Residuals:
##                     1                     2                     3
##   0.000000000000030395 -62.525834628775456281  62.525834628774937585
##                     4                     5                     6
##   6.100081427199171458  -0.000000000000657321  -6.100081427198049688
##                     7                     8                     9
##  -0.000000000000016500  -0.000000000000030711  -0.000000000000037817
##                    10                    11                    12
##  -0.000000000000046254  -0.000000000000551628   0.000000000000559039
##                    13                    14                    15
##   0.000000000000342324  -0.000000000000350455  -0.000000000000027159
##                    16                    17                    18
```

```
##  62.525834628776095769   0.000000000000946285 -62.525834628776905788
##                     19                     20                     21
##  -0.000000000000492564   0.000000000001017339  -0.000000000000925995
##                     22                     23                     24
##  -0.000000000000020053   0.000000000000075870  -0.000000000000048475
##                     25                     26                     27
##   0.000000000000054554   0.000000000000072317  62.525834628777573698
##                     28                     29                     30
##  -0.000000000000716385  -0.000000000001277714 -62.525834628775569968
##                     31                     32                     33
##  -0.000000000000223446 -62.525834628775676549  62.525834628776330248
##                     34                     35                     36
##  -0.000000000000454372 -99.126323191960651116  99.126323191961333237
##                     37                     38                     39
##  -0.000000000000987279   0.000000000000142484   0.000000000000394726
##                     40                     41                     42
##   0.000000000000441800  -0.000000000000227443  -0.000000000000035596
##                     43                     44                     45
##   0.000000000000321007  -6.100081427199207873   0.000000000000138043
##                     46                     47                     48
##   6.100081427198482231  99.126323191960423742 -99.126323191961347447
##                     49                     50                     51
##   0.000000000000967601 -62.525834628775598389  62.525834628775591284
##                     52                     53                     54
##  -0.000000000000891800  -0.000000000000003622   0.000000000000015030
##                     55                     56                     57
##   0.000000000000012365   0.000000000000000375  -0.000000000000307823
##                     58                     59                     60
##   0.000000000000404496  -0.000000000000074676  -0.000000000000026714
##                     61                     62                     63
##  -0.000000000000007175  -0.000000000000010727   0.000000000000062103
##                     64                     65                     66
##  -0.000000000000021385   0.000000000000905873  -0.000000000000003622
##                     67                     68                     69
##  99.126323191961276393 -99.126323191961859038   0.000000000000446685
##                     70                     71                     72
##  -0.000000000000344238  -0.000000000000021829   0.000000000000579467
##                     73                     74                     75
##   0.000000000000995135  -0.000000000000971292   0.000000000000174458
##                     76                     77                     78
## -62.525834628776010504   0.000000000000105180  62.525834628775662338
##                     79                     80                     81
## -99.126323191961660086  99.126323191962100623  -0.000000000000430392
##
## Coefficients: (12 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      356660.9    10613.9  33.603   0.0189 *
## monthMAY                         -15803.4      976.2 -16.189   0.0393 *
## monthJUN                        -244958.4     6892.4 -35.540   0.0179 *
## monthJUL                          -8746.9     1074.9  -8.137   0.0778 .
## monthAUG                         -96512.7     3202.2 -30.139   0.0211 *
## monthSEP                         -45331.9     1628.7 -27.834   0.0229 *
## monthOCT                         -12070.1     2100.8  -5.745   0.1097
## day2                              35637.5     1083.4  32.895   0.0193 *
```

```
## day3                                  33956.9     1346.8  25.213   0.0252 *
## day4                                  21814.0     2294.5   9.507   0.0667 .
## day5                                  55308.9     1489.7  37.128   0.0171 *
## day6                                  24819.3     1645.0  15.088   0.0421 *
## day7                                  23651.6     1558.7  15.174   0.0419 *
## day8                                  31170.6     1722.3  18.098   0.0351 *
## day9                                  11246.9     1757.6   6.399   0.0987 .
## day10                                -98298.8     4055.6 -24.238   0.0263 *
## day11                                -26822.6     1458.7 -18.388   0.0346 *
## day12                                -70796.7     2468.6 -28.679   0.0222 *
## day13                                -19444.6     1569.4 -12.390   0.0513 .
## day14                                 -5794.9      571.2 -10.145   0.0625 .
## day15                                 -9980.3      680.9 -14.658   0.0434 *
## day16                                 26585.6     1335.0  19.915   0.0319 *
## day17                                 70581.9     1997.1  35.343   0.0180 *
## day18                                -23119.0     1076.1 -21.483   0.0296 *
## day19                                -30928.0     1590.1 -19.450   0.0327 *
## day20                                 88628.5     3323.5  26.667   0.0239 *
## day21                                 85592.5     3313.6  25.830   0.0246 *
## day22                                 95911.8     3479.2  27.568   0.0231 *
## day23                               -110216.7     4108.0 -26.830   0.0237 *
## day24                                -68160.0     3355.4 -20.313   0.0313 *
## day25                               -111051.8     3915.3 -28.364   0.0224 *
## day26                                -73022.2     2899.3 -25.186   0.0253 *
## day27                                -74159.4     2916.4 -25.428   0.0250 *
## day28                                 -5717.8     1202.5  -4.755   0.1320
## day29                                -15723.3     1214.8 -12.943   0.0491 *
## day30                                -11443.6      846.3 -13.522   0.0470 *
## day31                                -52565.1     1858.8 -28.279   0.0225 *
## day_of_weekTuesday                   -26299.4     1001.4 -26.264   0.0242 *
## day_of_weekWednesday                  -2957.3      469.3  -6.302   0.1002
## day_of_weekThursday                   39490.7     1276.4  30.939   0.0206 *
## day_of_weekFriday                     24128.7     2916.4   8.273   0.0766 .
## day_of_weekSaturday                   26417.0      572.5  46.143   0.0138 *
## day_of_weekSunday                    -96727.5     3255.4 -29.713   0.0214 *
## opponentAstros                      -176158.8     4588.3 -38.393   0.0166 *
## opponentBraves                      -138535.7     3649.5 -37.960   0.0168 *
## opponentBrewers                     -258839.9     7032.3 -36.807   0.0173 *
## opponentCardinals                   -251094.4     7213.0 -34.811   0.0183 *
## opponentCubs                        -193423.5     5543.9 -34.889   0.0182 *
## opponentGiants                      -274927.0     7905.5 -34.777   0.0183 *
## opponentMarlins                      -93121.7     2749.1 -33.873   0.0188 *
## opponentMets                        -126745.6     3668.4 -34.550   0.0184 *
## opponentNationals                   -251811.9     6834.1 -36.846   0.0173 *
## opponentPadres                      -245440.2     7147.0 -34.342   0.0185 *
## opponentPhillies                    -296491.6     8751.2 -33.880   0.0188 *
## opponentPirates                     -211795.7     5929.7 -35.718   0.0178 *
## opponentReds                        -306950.7     8927.7 -34.382   0.0185 *
## opponentRockies                     -214993.1     5888.4 -36.511   0.0174 *
## opponentSnakes                      -264881.5     7415.9 -35.718   0.0178 *
## opponentWhite Sox                    -75407.6     2401.5 -31.400   0.0203 *
## temp                                   1880.7      121.2  15.512   0.0410 *
## pmax(0, temp - 23)                    -1537.0      130.5 -11.774   0.0539 .
## day_nightNight                       -81602.0     2659.5 -30.683   0.0207 *
```

```
## capYES                             312924.2    10981.8  28.495   0.0223 *
## shirtYES                           -13210.6     9516.3  -1.388   0.3974  
## bobbleheadYES                      -31079.2     1345.8 -23.094   0.0275 *
## skiesCloudy                        -30021.3     1162.0 -25.836   0.0246 *
## fireworksYES                        35857.8     2400.3  14.939   0.0426 *
## day_of_weekTuesday:bobbleheadYES    80567.6     2510.9  32.088   0.0198 *
## day_of_weekWednesday:bobbleheadYES       NA         NA      NA       NA  
## day_of_weekThursday:bobbleheadYES   99358.0     3179.3  31.251   0.0204 *
## day_of_weekFriday:bobbleheadYES          NA         NA      NA       NA  
## day_of_weekSaturday:bobbleheadYES        NA         NA      NA       NA  
## day_of_weekSunday:bobbleheadYES          NA         NA      NA       NA  
## bobbleheadYES:skiesCloudy           67137.1     2057.0  32.638   0.0195 *
## temp:capYES                        -12168.2      450.9 -26.984   0.0236 *
## temp:shirtYES                        1143.9      445.3   2.569   0.2364  
## opponentAstros:fireworksYES          6062.7     1304.0   4.649   0.1349  
## opponentBraves:fireworksYES              NA         NA      NA       NA  
## opponentBrewers:fireworksYES             NA         NA      NA       NA  
## opponentCardinals:fireworksYES     -37974.8     1536.3 -24.719   0.0257 *
## opponentCubs:fireworksYES          -78784.6     3400.9 -23.166   0.0275 *
## opponentGiants:fireworksYES              NA         NA      NA       NA  
## opponentMarlins:fireworksYES       -78820.9     2371.4 -33.238   0.0191 *
## opponentMets:fireworksYES           60769.9     2256.0  26.937   0.0236 *
## opponentNationals:fireworksYES           NA         NA      NA       NA  
## opponentPadres:fireworksYES        -31112.5     1583.2 -19.652   0.0324 *
## opponentPhillies:fireworksYES            NA         NA      NA       NA  
## opponentPirates:fireworksYES             NA         NA      NA       NA  
## opponentReds:fireworksYES                NA         NA      NA       NA  
## opponentRockies:fireworksYES       -75750.1     2368.1 -31.988   0.0199 *
## opponentSnakes:fireworksYES         72592.9     3101.5  23.406   0.0272 *
## opponentWhite Sox:fireworksYES           NA         NA      NA       NA  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354.5 on 1 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:  0.9982
## F-statistic: 554.8 on 79 and 1 DF,  p-value: 0.03376
```

It gives an adjusted R2 value of 0.9981747. At first glance, it looks highly promising, but it is deceiving. When we discard variable **day** from the model, the AIC sharply increase and adjusted R2 decrease.

```
AIC(update(model_best, .~. - day), model_best)
```

```
##                                    df       AIC
## update(model_best, . ~ . - day) 52 1658.9498
## model_best                         81  986.9728
```

```
summary(update(model_best, .~. - day))$adj.r.squared
```

```
## [1] 0.5009968
```

We examined the effect of day on attendance. It seems highly unrelated, but probably by chance, it has a huge effect on attendance in this model. So, it is possible that the step function results in an overfitted model. We can apply **cross-validation** using the library *caret* to analyze if it overfits to train data.

```
# First set the random seed since cross-validation randomly assigns rows to each
# fold and we want to be able to produce our model exactly.
set.seed(42)
```

```r
model <- train(  attend ~ month+
                             day+
                             day_of_week+
                             opponent+
                             temp+
                             pmax(0, temp - 23)+
                             day_night+
                             cap+
                             shirt+
                             bobblehead+
                             bobblehead*day_of_week+
                             skies*fireworks+
                             skies*day_night*fireworks+
                             skies*day_night*bobblehead+
                             cap*shirt*temp+
                             opponent*fireworks+
                             fireworks*day_night*skies, events,  method ="lm",
        trControl = trainControl(method ="cv", number =10, verboseIter =TRUE))

print(model)
```

The cross-validation result supports our hypothesis. The model is highly poor as the R2 value is 0.09.

We are going to remove the variable "day" from the step function and try again. Also, in this case instead of backward, the algorithm will move in a forward manner.

```r
# Fit the full model with all possible predictors, including interaction terms
events[, day := factor(day)]

model_full <- lm(attend ~ (month+
                             day_of_week+
                             opponent+
                             temp+
                             pmax(0, temp - 23)+
                             day_night+
                             skies+
                             cap+
                             shirt+
                             bobblehead+
                             skies*fireworks+
                             day_night*fireworks+
                             shirt*temp+
                             opponent*fireworks+
                             fireworks*opponent), data = events)

# Use stepwise selection to find the best model with highest adjusted R-squared, including interaction
model_best <- step(model_full, direction = "forward", k = log(nrow(events)))
```

```
## Start:  AIC=1543.61
## attend ~ (month + day_of_week + opponent + temp + pmax(0, temp -
##     23) + day_night + skies + cap + shirt + bobblehead + skies *
##     fireworks + day_night * fireworks + shirt * temp + opponent *
##     fireworks + fireworks * opponent)
```

The result has some useful insights. The result suggests that opponents do not have an effect on attendance and the effect of fireworks is nearly negligible.

Model Development

We see that temperature, bobblehead and some months & days have effects on attendance. Also, we stated that although we have small observations, temperature and shirt have an association. We can start our model by these variables.

```
model1 <- lm(attend ~ temp + pmax(0, temp - 23) + bobblehead  + month + day_of_week + temp*shirt, data =

summary(model1)
```

```
##
## Call:
## lm(formula = attend ~ temp + pmax(0, temp - 23) + bobblehead +
##     month + day_of_week + temp * shirt, data = events)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10977.8  -3236.0    -37.8   1813.3  14156.1
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           19445.74    7063.78   2.753 0.007708 **
## temp                    774.34     394.68   1.962 0.054193 .
## pmax(0, temp - 23)    -1434.06     603.19  -2.377 0.020480 *
## bobbleheadYES          9864.99    2398.18   4.114 0.000115 ***
## monthMAY              -4048.19    2499.70  -1.619 0.110343
## monthJUN               3650.97    3011.75   1.212 0.229946
## monthJUL                411.05    3083.04   0.133 0.894360
## monthAUG                143.38    3326.50   0.043 0.965757
## monthSEP                 87.16    4200.50   0.021 0.983511
## monthOCT               -826.75    5134.62  -0.161 0.872596
## day_of_weekTuesday     8743.88    2699.50   3.239 0.001916 **
## day_of_weekWednesday   3296.01    2449.45   1.346 0.183250
## day_of_weekThursday    2130.24    3370.89   0.632 0.529705
## day_of_weekFriday      6037.01    2487.58   2.427 0.018102 *
## day_of_weekSaturday    7764.35    2491.10   3.117 0.002753 **
## day_of_weekSunday      7140.55    2529.81   2.823 0.006368 **
## shirtYES              63559.70   35291.68   1.801 0.076492 .
## temp:shirtYES         -2577.32    1527.75  -1.687 0.096549 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5817 on 63 degrees of freedom
## Multiple R-squared:  0.6129, Adjusted R-squared:  0.5085
## F-statistic: 5.868 on 17 and 63 DF,  p-value: 0.00000009647
```

We are doing multiple linear regression and there are some conditions associated with linear regression, ie., linearity, nearly normal residuals, and constant variability. Using diagnostic tools we will assess whether these conditions have been met or not.

Let's assess the model results and diagnostics.

The p-values of variables temp, pmax(0, temp - 23), bobbleheadYES, day_of_weekTuesday, day_of_weekFriday, day_of_weekSaturday & day_of_weekSunday are < 0.05, so they are statistically significant. Also, p-value of the model is less than 0.05 and it indicates that there is evidence against the null hypothesis and that the relationship between the dependent variable and at least one independent variable in the model is statistically significant.

The adjusted R-squared of 0.51 indicates that the model explains approximately 51% of the variation in attendance, which means that there is still a substantial amount of unexplained variability in the data.

It is seen that not any of the month has a p-value that is less than 0.05. Let's check if **month** variable improves the model or not.

```
AIC(update(model1, .~. - month), model1)
```

```
##                                 df      AIC
## update(model1, . ~ . - month) 13 1652.590
## model1                        19 1651.825
```

AIC says that months' effect is negligible. We can conduct **F-test** also.

```
anova(update(model1, .~. - month), model1)
```

```
## Analysis of Variance Table
##
## Model 1: attend ~ temp + pmax(0, temp - 23) + bobblehead + day_of_week +
##     shirt + temp:shirt
## Model 2: attend ~ temp + pmax(0, temp - 23) + bobblehead + month + day_of_week +
##     temp * shirt
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1     69 2495940234
## 2     63 2132031288  6 363908946 1.7922 0.1151
```

Null is that small model is correct. The null is consistent with data since p-value is large. Hence, month variable is not important. It may be the case that day_of_week and month are associated as it seems logical to expect different effects from days in different months. So, we can add an interaction term.

```
model2 <- lm(attend ~ temp + pmax(0, temp - 23) + bobblehead  + month*day_of_week + temp*shirt, data = e

summary(model2)
```

```
##
## Call:
## lm(formula = attend ~ temp + pmax(0, temp - 23) + bobblehead +
##     month * day_of_week + temp * shirt, data = events)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10217  -1903      0   1691  10758
##
## Coefficients: (6 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      16557.9    10097.8   1.640  0.11056
## temp                               613.6      530.5   1.157  0.25571
## pmax(0, temp - 23)                -841.3      836.5  -1.006  0.32187
## bobbleheadYES                    10995.9     3741.8   2.939  0.00598 **
## monthMAY                          6311.9     6721.2   0.939  0.35450
## monthJUN                         14402.6    10952.3   1.315  0.19756
## monthJUL                          3859.5     6850.3   0.563  0.57696
## monthAUG                          4894.0     7777.8   0.629  0.53354
## monthSEP                          4234.7     9038.2   0.469  0.64248
```

```
## monthOCT                      4546.4     9237.6   0.492  0.62586
## day_of_weekTuesday           22403.7     6782.5   3.303  0.00231 **
## day_of_weekWednesday          1661.0     6699.0   0.248  0.80571
## day_of_weekThursday           3179.3     7807.8   0.407  0.68650
## day_of_weekFriday            12134.8     6704.2   1.810  0.07941 .
## day_of_weekSaturday          17294.3     6954.0   2.487  0.01812 *
## day_of_weekSunday            10755.7     7807.8   1.378  0.17761
## shirtYES                      7325.8     8177.5   0.896  0.37682
## monthMAY:day_of_weekTuesday   -22561.7   8435.2  -2.675  0.01155 *
## monthJUN:day_of_weekTuesday   -20740.1  13318.6  -1.557  0.12895
## monthJUL:day_of_weekTuesday   -12967.9   8215.7  -1.578  0.12401
## monthAUG:day_of_weekTuesday   -12883.8   9593.7  -1.343  0.18846
## monthSEP:day_of_weekTuesday   -23333.6  13394.4  -1.742  0.09082 .
## monthOCT:day_of_weekTuesday   -14010.1  10334.1  -1.356  0.18439
## monthMAY:day_of_weekWednesday  -7972.9   8378.0  -0.952  0.34819
## monthJUN:day_of_weekWednesday   -786.5  13001.9  -0.060  0.95213
## monthJUL:day_of_weekWednesday  11639.6   8439.4   1.379  0.17711
## monthAUG:day_of_weekWednesday   1483.6   8355.6   0.178  0.86016
## monthSEP:day_of_weekWednesday  14448.3  10486.7   1.378  0.17754
## monthOCT:day_of_weekWednesday  -1726.4  10296.9  -0.168  0.86787
## monthMAY:day_of_weekThursday  -12162.4  10086.0  -1.206  0.23644
## monthJUN:day_of_weekThursday  -10015.5  14489.7  -0.691  0.49427
## monthJUL:day_of_weekThursday        NA       NA      NA       NA
## monthAUG:day_of_weekThursday    5791.2  10849.0   0.534  0.59706
## monthSEP:day_of_weekThursday    6134.4  11066.6   0.554  0.58310
## monthOCT:day_of_weekThursday        NA       NA      NA       NA
## monthMAY:day_of_weekFriday     -7843.0   8190.3  -0.958  0.34523
## monthJUN:day_of_weekFriday    -10577.3  12638.0  -0.837  0.40865
## monthJUL:day_of_weekFriday     -2565.1   9307.5  -0.276  0.78458
## monthAUG:day_of_weekFriday     -7022.5   8514.5  -0.825  0.41542
## monthSEP:day_of_weekFriday     -7480.2   9551.3  -0.783  0.43912
## monthOCT:day_of_weekFriday          NA       NA      NA       NA
## monthMAY:day_of_weekSaturday  -14445.3   8472.3  -1.705  0.09759 .
## monthJUN:day_of_weekSaturday  -16392.9  12698.7  -1.291  0.20571
## monthJUL:day_of_weekSaturday   -8579.4   9421.0  -0.911  0.36907
## monthAUG:day_of_weekSaturday   -8810.1   9096.9  -0.968  0.33986
## monthSEP:day_of_weekSaturday  -10809.1   9426.4  -1.147  0.25976
## monthOCT:day_of_weekSaturday        NA       NA      NA       NA
## monthMAY:day_of_weekSunday     -4624.1   8949.4  -0.517  0.60881
## monthJUN:day_of_weekSunday     -2325.8  15794.7  -0.147  0.88383
## monthJUL:day_of_weekSunday     -2906.1   9447.0  -0.308  0.76031
## monthAUG:day_of_weekSunday     -3095.6   9559.6  -0.324  0.74812
## monthSEP:day_of_weekSunday     -9214.1  10251.9  -0.899  0.37528
## monthOCT:day_of_weekSunday          NA       NA      NA       NA
## temp:shirtYES                       NA       NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5470 on 33 degrees of freedom
## Multiple R-squared:  0.8208, Adjusted R-squared:  0.5655
## F-statistic: 3.215 on 47 and 33 DF,  p-value: 0.0003443
```

Use AIC to see if the interaction term is important.

```
AIC(update(model2, .~. - month:day_of_week), model2)
```

```
##                                          df      AIC
## update(model2, . ~ . - month:day_of_week) 19 1651.825
## model2                                   49 1649.465
```

As seen although degrees of freedom increases a lot, AIC decreases. So, the interaction term is important. However, still adjusted R2 value is small. Probably there are other variables and interaction terms that have an effect on attendance. Now, we will add new variables to the model.

It is reasonable to think that fans want to create huge shows in some matches to impress their biggest opponents. So, in these matches, the fireworks may be the part of the show and it can increase attendance. The variable **fireworks:opponent** might reflect this interpretation.

```
model3 <- lm(attend ~ temp + pmax(0, temp - 23) + bobblehead  + month*day_of_week + temp*shirt + firewo

summary(model3)
```

```
##
## Call:
## lm(formula = attend ~ temp + pmax(0, temp - 23) + bobblehead +
##     month * day_of_week + temp * shirt + fireworks * opponent,
##     data = events)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6402.9  -265.6     0.0   101.4  6402.9
##
## Coefficients: (20 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -9387.8    28565.3  -0.329  0.74729
## temp                           1256.6     1091.6   1.151  0.26895
## pmax(0, temp - 23)            -1249.2     1287.2  -0.970  0.34827
## bobbleheadYES                 19959.2     5587.3   3.572  0.00306 **
## monthMAY                      11822.5    21253.6   0.556  0.58681
## monthJUN                       8385.1    15208.0   0.551  0.59008
## monthJUL                       7112.9    22769.8   0.312  0.75936
## monthAUG                       5028.8    20994.3   0.240  0.81416
## monthSEP                      -1493.4    14787.1  -0.101  0.92099
## monthOCT                       1977.8    23448.7   0.084  0.93398
## day_of_weekTuesday            16647.0     7056.2   2.359  0.03337 *
## day_of_weekWednesday          -2809.8     6530.0  -0.430  0.67354
## day_of_weekThursday           -4476.4     8589.5  -0.521  0.61041
## day_of_weekFriday            -15659.5    15605.6  -1.003  0.33267
## day_of_weekSaturday           22868.9    12495.3   1.830  0.08859 .
## day_of_weekSunday              9652.6     9753.7   0.990  0.33915
## shirtYES                      26430.0    12205.6   2.165  0.04812 *
## fireworksYES                  26093.7     9511.4   2.743  0.01585 *
## opponentAstros                 3733.5    26954.4   0.139  0.89181
## opponentBraves                15658.5    21523.3   0.728  0.47890
## opponentBrewers                6578.4    26729.7   0.246  0.80917
## opponentCardinals              7998.6    26571.3   0.301  0.76782
## opponentCubs                   9065.7    32364.1   0.280  0.78348
## opponentGiants                12081.1    26830.0   0.450  0.65940
## opponentMarlins                6880.0    31330.3   0.220  0.82936
```

```
## opponentMets                       11167.8    17391.9   0.642  0.53116
## opponentNationals                  -6843.2    19867.5  -0.344  0.73563
## opponentPadres                     15475.7    20662.3   0.749  0.46627
## opponentPhillies                   17040.7    28939.4   0.589  0.56535
## opponentPirates                    24600.1    22206.6   1.108  0.28663
## opponentReds                        7104.7    28745.7   0.247  0.80838
## opponentRockies                     8319.5    26450.1   0.315  0.75775
## opponentSnakes                      3425.7    26982.6   0.127  0.90078
## opponentWhite Sox                  15952.6    16303.2   0.978  0.34443
## monthMAY:day_of_weekTuesday       -24144.8     8462.5  -2.853  0.01277 *
## monthJUN:day_of_weekTuesday        -4199.7    15422.8  -0.272  0.78936
## monthJUL:day_of_weekTuesday       -10706.0     7996.5  -1.339  0.20196
## monthAUG:day_of_weekTuesday       -15855.4     9841.0  -1.611  0.12945
## monthSEP:day_of_weekTuesday       -35976.0    14970.6  -2.403  0.03069 *
## monthOCT:day_of_weekTuesday        -7783.3     9997.5  -0.779  0.44922
## monthMAY:day_of_weekWednesday      -6220.3     7989.2  -0.779  0.44918
## monthJUN:day_of_weekWednesday      23431.4    16070.0   1.458  0.16688
## monthJUL:day_of_weekWednesday       -931.8    10132.2  -0.092  0.92803
## monthAUG:day_of_weekWednesday       8251.7     7985.0   1.033  0.31894
## monthSEP:day_of_weekWednesday      19859.2     9914.2   2.003  0.06492 .
## monthOCT:day_of_weekWednesday       3214.5     9635.5   0.334  0.74362
## monthMAY:day_of_weekThursday       -4152.0    10551.3  -0.394  0.69987
## monthJUN:day_of_weekThursday       -5550.8    16700.6  -0.332  0.74454
## monthJUL:day_of_weekThursday            NA         NA      NA       NA
## monthAUG:day_of_weekThursday       11140.6    12144.4   0.917  0.37450
## monthSEP:day_of_weekThursday       21737.2    14148.7   1.536  0.14674
## monthOCT:day_of_weekThursday            NA         NA      NA       NA
## monthMAY:day_of_weekFriday         -7638.2    14866.6  -0.514  0.61542
## monthJUN:day_of_weekFriday         -8827.2    10077.2  -0.876  0.39584
## monthJUL:day_of_weekFriday         -8670.6    22998.2  -0.377  0.71182
## monthAUG:day_of_weekFriday          -795.0    11629.3  -0.068  0.94647
## monthSEP:day_of_weekFriday          -233.1    12059.0  -0.019  0.98485
## monthOCT:day_of_weekFriday              NA         NA      NA       NA
## monthMAY:day_of_weekSaturday      -17627.5    13874.3  -1.271  0.22461
## monthJUN:day_of_weekSaturday      -17740.6     9898.3  -1.792  0.09472 .
## monthJUL:day_of_weekSaturday      -30923.7    23519.1  -1.315  0.20970
## monthAUG:day_of_weekSaturday      -10691.7     9626.3  -1.111  0.28542
## monthSEP:day_of_weekSaturday       -7802.6     9246.1  -0.844  0.41292
## monthOCT:day_of_weekSaturday            NA         NA      NA       NA
## monthMAY:day_of_weekSunday         -3857.1    11848.8  -0.326  0.74960
## monthJUN:day_of_weekSunday              NA         NA      NA       NA
## monthJUL:day_of_weekSunday        -12054.5    21395.8  -0.563  0.58207
## monthAUG:day_of_weekSunday              NA         NA      NA       NA
## monthSEP:day_of_weekSunday              NA         NA      NA       NA
## monthOCT:day_of_weekSunday              NA         NA      NA       NA
## temp:shirtYES                           NA         NA      NA       NA
## fireworksYES:opponentAstros         8470.2     8174.8   1.036  0.31771
## fireworksYES:opponentBraves             NA         NA      NA       NA
## fireworksYES:opponentBrewers            NA         NA      NA       NA
## fireworksYES:opponentCardinals      4480.7     6055.8   0.740  0.47158
## fireworksYES:opponentCubs            289.7     8526.5   0.034  0.97338
## fireworksYES:opponentGiants             NA         NA      NA       NA
## fireworksYES:opponentMarlins            NA         NA      NA       NA
## fireworksYES:opponentMets          10346.1    10043.8   1.030  0.32044
```

```
## fireworksYES:opponentNationals  26728.8    11515.7   2.321  0.03589 *
## fireworksYES:opponentPadres           NA         NA      NA       NA
## fireworksYES:opponentPhillies         NA         NA      NA       NA
## fireworksYES:opponentPirates          NA         NA      NA       NA
## fireworksYES:opponentReds             NA         NA      NA       NA
## fireworksYES:opponentRockies          NA         NA      NA       NA
## fireworksYES:opponentSnakes           NA         NA      NA       NA
## fireworksYES:opponentWhite Sox        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4919 on 14 degrees of freedom
## Multiple R-squared:  0.9385, Adjusted R-squared:  0.6486
## F-statistic: 3.237 on 66 and 14 DF,  p-value: 0.008967
```

The adjusted R2 is increased to 0.65 from 0.57 by the new interaction term - **fireworks:opponent**.

Let's check if the model is improved by AIC.

```
AIC(model2, model3)
```

```
##        df      AIC
## model2 49 1649.465
## model3 68 1600.819
```

AIC suggests that the model is improved. So, with the new variable, the model is able to work better in an unseen data. Although we made an important improvement, it is possible that there are other variables having effect on attendance. We did not add one promotion to our model which is **cap**. Let's add it to the model.

```
model4 <- lm(attend ~ temp + pmax(0, temp - 23) + bobblehead  + month*day_of_week + temp*shirt + firewo

summary(model4)
```

```
##
## Call:
## lm(formula = attend ~ temp + pmax(0, temp - 23) + bobblehead +
##     month * day_of_week + temp * shirt + fireworks * opponent +
##     cap, data = events)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5746.5  -329.1     0.0   229.3  5806.5
##
## Coefficients: (20 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1704.3    29615.9   0.058  0.95498
## temp                                820.3     1135.0   0.723  0.48268
## pmax(0, temp - 23)                 -805.5     1320.7  -0.610  0.55245
## bobbleheadYES                     17700.5     5816.5   3.043  0.00942 **
## monthMAY                          16760.0    21332.6   0.786  0.44616
## monthJUN                           9587.2    15010.5   0.639  0.53411
## monthJUL                          12701.0    22903.9   0.555  0.58863
## monthAUG                          11544.0    21378.6   0.540  0.59834
## monthSEP                           2724.7    14982.1   0.182  0.85849
## monthOCT                           8474.9    23720.7   0.357  0.72662
## day_of_weekTuesday                17737.9     7008.4   2.531  0.02508 *
```

```
## day_of_weekWednesday           -2591.6     6433.4  -0.403  0.69361
## day_of_weekThursday            -4912.7     8466.9  -0.580  0.57168
## day_of_weekFriday             -10787.2    15897.6  -0.679  0.50933
## day_of_weekSaturday            23204.9    12308.7   1.885  0.08194 .
## day_of_weekSunday              11733.8     9761.3   1.202  0.25078
## shirtYES                       22862.3    12383.6   1.846  0.08776 .
## fireworksYES                   20684.5    10398.3   1.989  0.06814 .
## opponentAstros                 -4456.5    27411.0  -0.163  0.87335
## opponentBraves                 11547.5    21472.3   0.538  0.59981
## opponentBrewers                  127.8    26868.7   0.005  0.99628
## opponentCardinals                728.2    26862.1   0.027  0.97879
## opponentCubs                    5790.4    31989.4   0.181  0.85915
## opponentGiants                  4475.2    27174.4   0.165  0.87172
## opponentMarlins                 -957.9    31540.4  -0.030  0.97623
## opponentMets                    8114.5    17316.2   0.469  0.64711
## opponentNationals              -6413.4    19568.9  -0.328  0.74833
## opponentPadres                 10156.2    20827.2   0.488  0.63392
## opponentPhillies                7189.7    29662.3   0.242  0.81226
## opponentPirates                20052.8    22196.2   0.903  0.38273
## opponentReds                    1819.9    28650.7   0.064  0.95032
## opponentRockies                  729.7    26807.6   0.027  0.97870
## opponentSnakes                 -3990.7    27284.3  -0.146  0.88596
## opponentWhite Sox              11612.7    16459.1   0.706  0.49292
## capYES                         -8259.9     6894.6  -1.198  0.25230
## monthMAY:day_of_weekTuesday   -22859.6     8402.7  -2.721  0.01750 *
## monthJUN:day_of_weekTuesday    -7035.8    15371.9  -0.458  0.65472
## monthJUL:day_of_weekTuesday    -8002.2     8192.0  -0.977  0.34648
## monthAUG:day_of_weekTuesday   -14680.1     9741.0  -1.507  0.15571
## monthSEP:day_of_weekTuesday   -33476.7    14890.0  -2.248  0.04254 *
## monthOCT:day_of_weekTuesday    -8859.3     9886.5  -0.896  0.38649
## monthMAY:day_of_weekWednesday  -6058.8     7869.0  -0.770  0.45509
## monthJUN:day_of_weekWednesday  19209.2    16213.5   1.185  0.25732
## monthJUL:day_of_weekWednesday   2026.5    10279.3   0.197  0.84676
## monthAUG:day_of_weekWednesday   7974.3     7867.1   1.014  0.32926
## monthSEP:day_of_weekWednesday  19670.8     9764.8   2.014  0.06513 .
## monthOCT:day_of_weekWednesday   3011.2     9490.6   0.317  0.75607
## monthMAY:day_of_weekThursday   -4131.8    10391.0  -0.398  0.69735
## monthJUN:day_of_weekThursday   -2068.6    16701.8  -0.124  0.90333
## monthJUL:day_of_weekThursday        NA         NA      NA       NA
## monthAUG:day_of_weekThursday   13650.5    12142.0   1.124  0.28124
## monthSEP:day_of_weekThursday   24139.4    14077.3   1.715  0.11011
## monthOCT:day_of_weekThursday        NA         NA      NA       NA
## monthMAY:day_of_weekFriday     -7680.0    14640.8  -0.525  0.60872
## monthJUN:day_of_weekFriday     -7954.5     9950.8  -0.799  0.43843
## monthJUL:day_of_weekFriday     -9466.6    22658.6  -0.418  0.68292
## monthAUG:day_of_weekFriday      -428.5    11456.7  -0.037  0.97073
## monthSEP:day_of_weekFriday      2596.3    12108.4   0.214  0.83354
## monthOCT:day_of_weekFriday          NA         NA      NA       NA
## monthMAY:day_of_weekSaturday  -18601.2    13687.7  -1.359  0.19727
## monthJUN:day_of_weekSaturday  -17304.3     9754.7  -1.774  0.09948 .
## monthJUL:day_of_weekSaturday  -30333.6    23167.1  -1.309  0.21309
## monthAUG:day_of_weekSaturday  -13479.1     9761.4  -1.381  0.19060
## monthSEP:day_of_weekSaturday   -6042.4     9223.4  -0.655  0.52382
## monthOCT:day_of_weekSaturday        NA         NA      NA       NA
```

```
## monthMAY:day_of_weekSunday      -4835.7    11697.4  -0.413  0.68605
## monthJUN:day_of_weekSunday           NA         NA      NA       NA
## monthJUL:day_of_weekSunday     -15475.9    21263.5  -0.728  0.47963
## monthAUG:day_of_weekSunday           NA         NA      NA       NA
## monthSEP:day_of_weekSunday           NA         NA      NA       NA
## monthOCT:day_of_weekSunday           NA         NA      NA       NA
## temp:shirtYES                        NA         NA      NA       NA
## fireworksYES:opponentAstros      7753.9     8072.8   0.960  0.35433
## fireworksYES:opponentBraves          NA         NA      NA       NA
## fireworksYES:opponentBrewers         NA         NA      NA       NA
## fireworksYES:opponentCardinals   4146.4     5970.3   0.695  0.49959
## fireworksYES:opponentCubs       -3836.5     9075.9  -0.423  0.67941
## fireworksYES:opponentGiants          NA         NA      NA       NA
## fireworksYES:opponentMarlins         NA         NA      NA       NA
## fireworksYES:opponentMets       10368.4     9891.3   1.048  0.31363
## fireworksYES:opponentNationals  24033.8    11561.7   2.079  0.05801 .
## fireworksYES:opponentPadres          NA         NA      NA       NA
## fireworksYES:opponentPhillies        NA         NA      NA       NA
## fireworksYES:opponentPirates         NA         NA      NA       NA
## fireworksYES:opponentReds            NA         NA      NA       NA
## fireworksYES:opponentRockies         NA         NA      NA       NA
## fireworksYES:opponentSnakes          NA         NA      NA       NA
## fireworksYES:opponentWhite Sox       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4844 on 13 degrees of freedom
## Multiple R-squared:  0.9446, Adjusted R-squared:  0.6592
## F-statistic: 3.309 on 67 and 13 DF,  p-value: 0.0102
```

Adjusted R2 is slightly increased from 0.6485638 to 0.6591604, but the p-value of capYES is more than 0.05. Let's check if it improves the model.

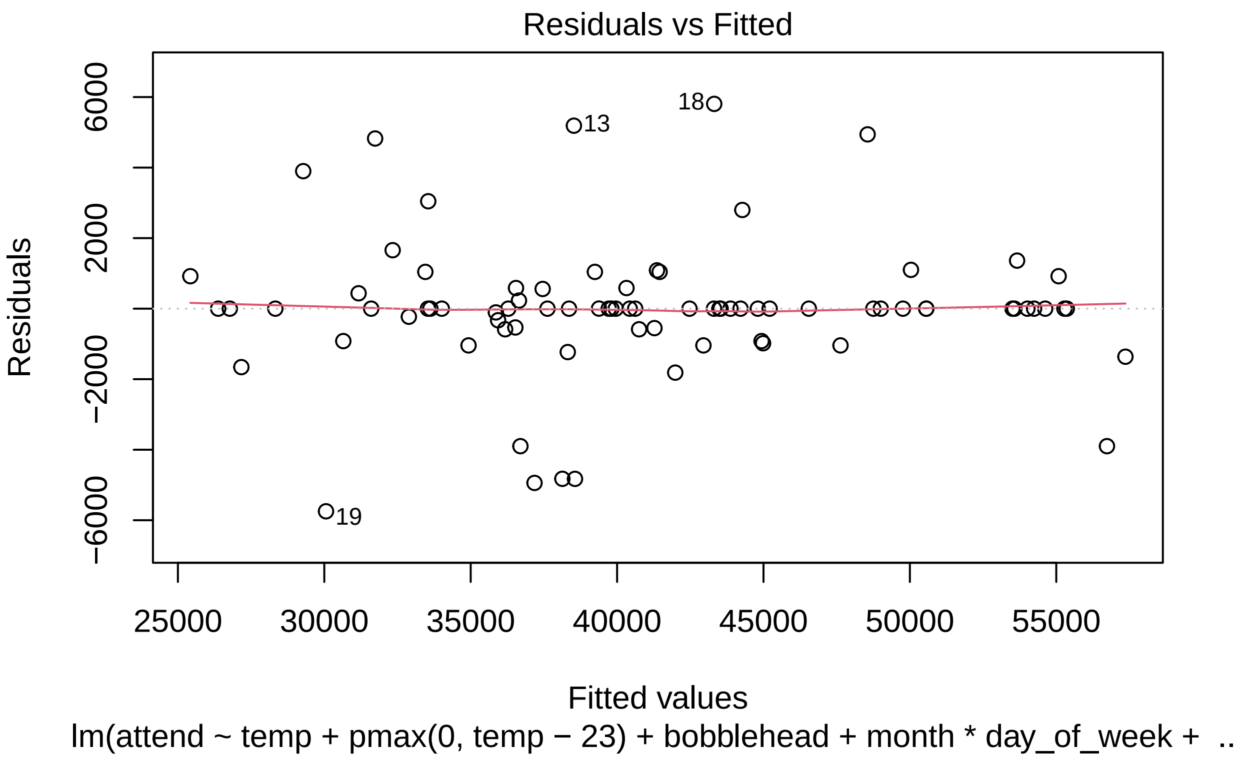
```
AIC(update(model4, .~. - cap), model4)
```

```
##                              df      AIC
## update(model4, . ~ . - cap) 68 1600.819
## model4                      69 1594.336
```

AIC suggests that it improves the model.

Now analyze diagnostic plots of model4.

```
plot(model4, which=1)
```



The ideal residual would be zero, because that would mean that the data point falls exactly on the regression line and that there is no difference between the **predicted** and **observed values** for that **particular data point**.

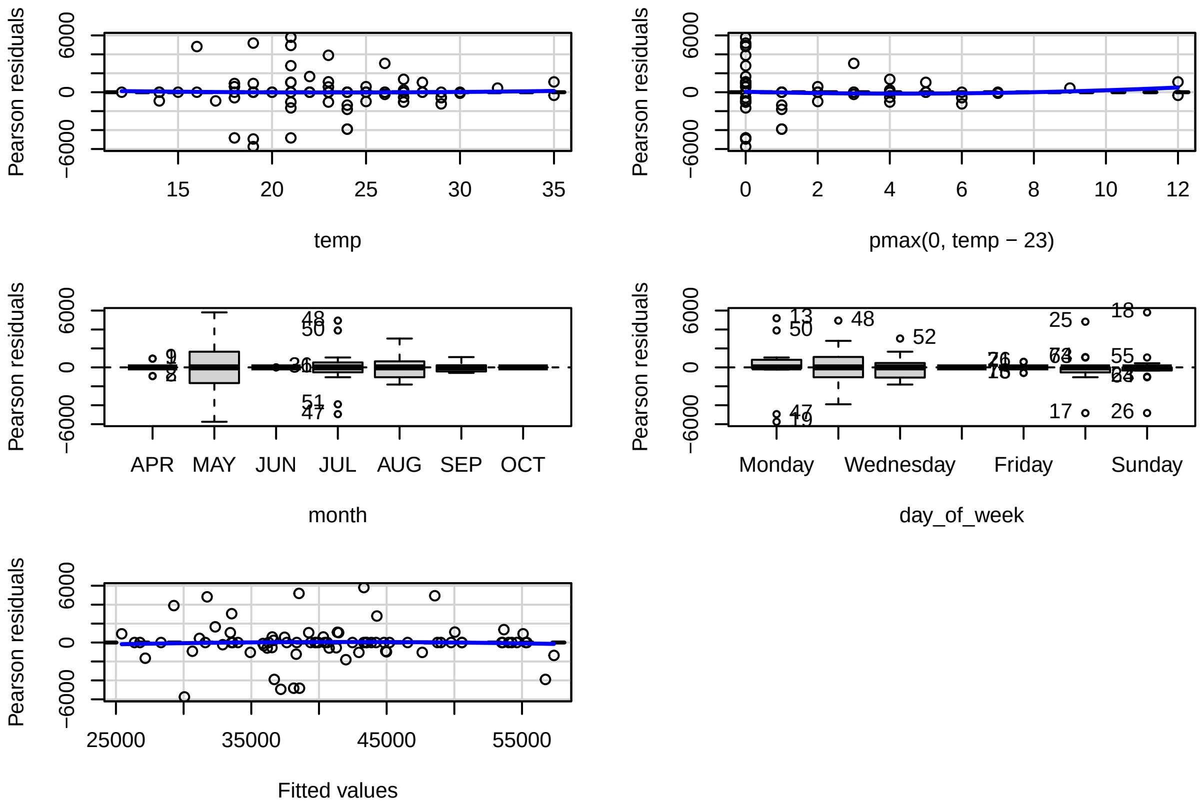
This is unlikely to happen, but we like **small residuals** and we want our **residuals** in the **residuals plot** to be **randomly scattered around zero.**

There's going to be some that are positive and some that are negative, because that corresponds to some points falling above the regression line, and other points falling below the regression line. And we want them to have absolutely no pattern, because **what we want is for the linear model is to capture all of the pattern in the data, and anything that's left over to be simply random scatter**.

The plot shows that residuals are centered at near zero. So, the model predicts very good. There is not any strong pattern in the residuals.

The residuals are not only be supposed to uncorrelated with fitted values, but also with each one of the predictor.

```
car::residualPlots(model4)
```




```
##                    Test stat Pr(>|Test stat|)
## temp                  0.3850           0.7069
## pmax(0, temp - 23)    0.4146           0.6857
## month
## day_of_week
## Tukey test           -0.3452           0.7300
```

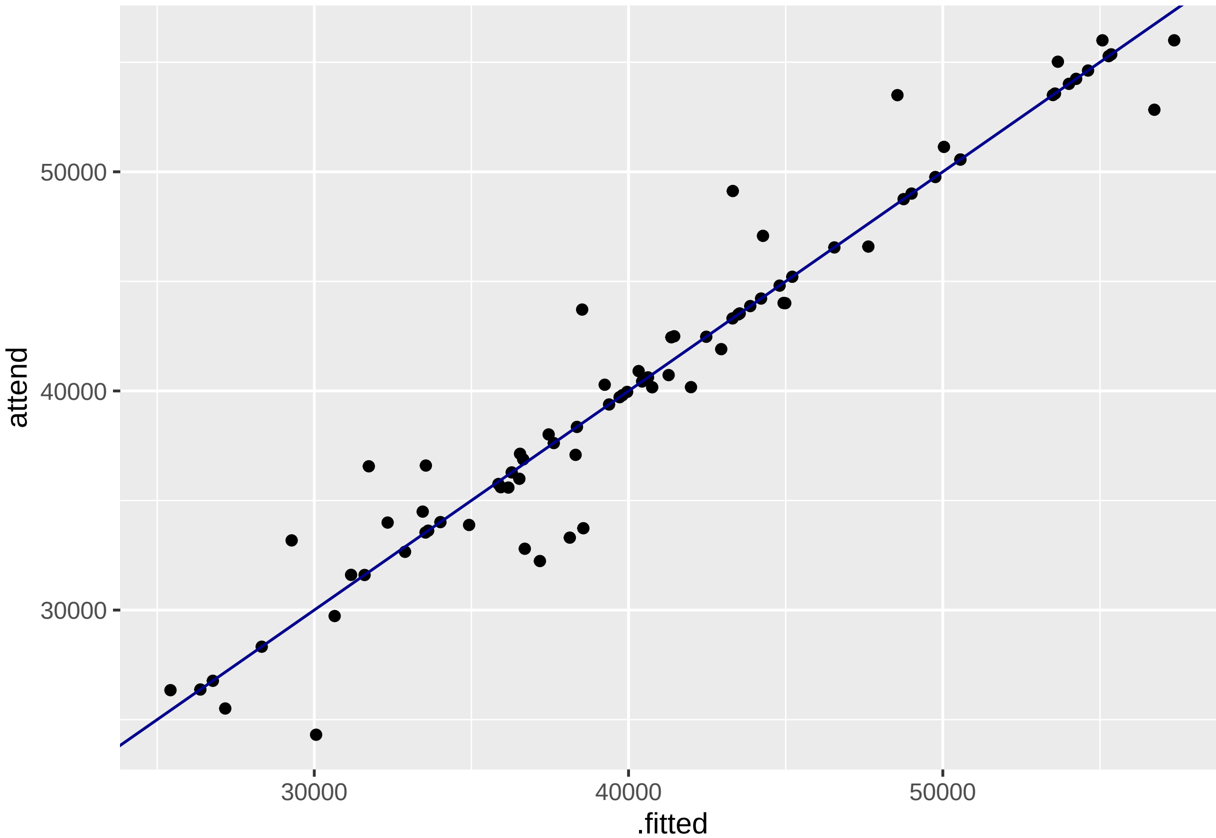
From the graph, we can conclude that there is not any systematic error in any of the variables as there is no pattern in any of them.

The next condition is **nearly normal residuals**, which says that residuals should be **nearly normally distributed, centered at zero**.

This condition may not be satisfied if there are unusual observations that don't follow the trend of the rest of the data.

We will check if the points are normally distributed around the **fitted line**.

```
# fitted vs attend
ggplot(model4, aes(.fitted, attend)) + geom_point() +
  geom_abline(color="darkblue")
```

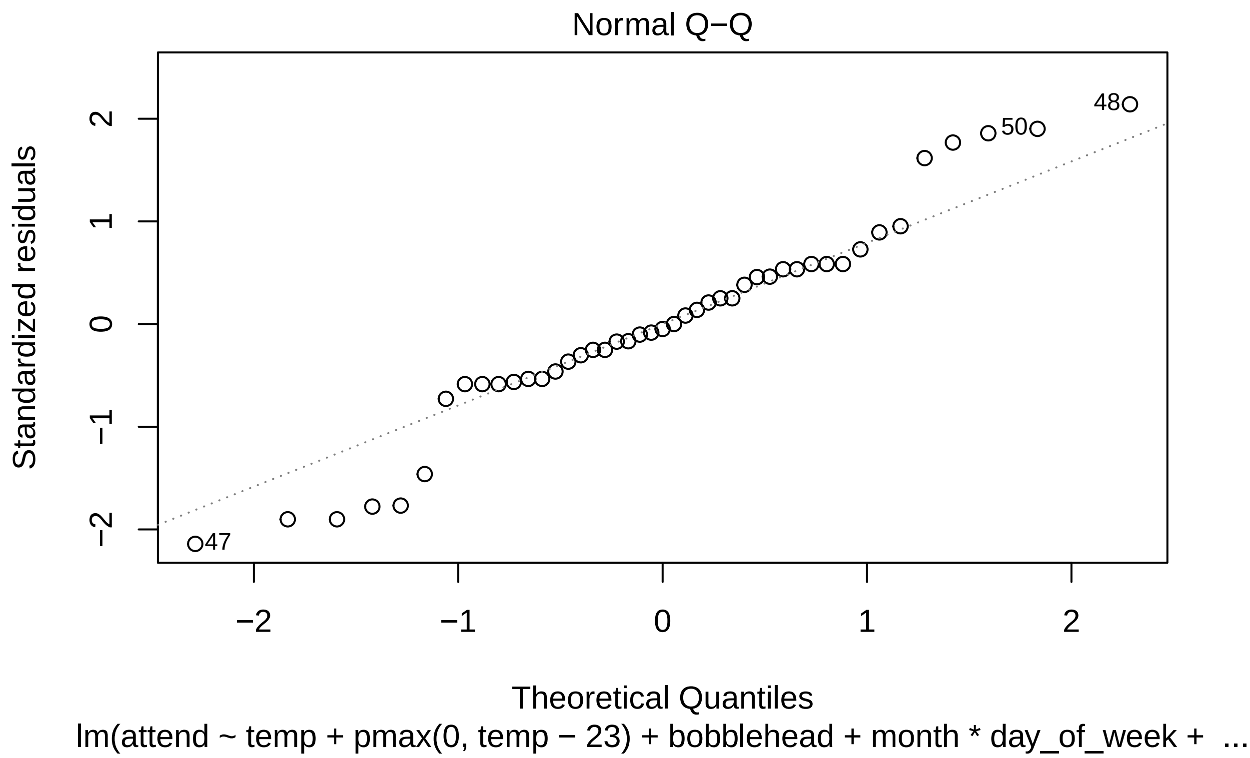


In this plot, the points are equally far from the line in both the upper and lower side of the line. So, the points are nearly normally distributed around the line.

The **Normal Q-Q plot** makes it even easier to check the normality.

```
plot(model4, which=2)
```

```
## Warning: not plotting observations with leverage one:
##   3, 4, 5, 6, 7, 10, 11, 12, 24, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 43, 44, 45, 46, 49, 53,
```



From the graph, although there are some deviations at tails, overall they fall along the diagonal line.

We can also apply **Shapiro Test** to see if these deviations are statistically significant.

```
shapiro.test(rstandard(model4)) # null says standardized residuals have normal distribution. Since p is
```

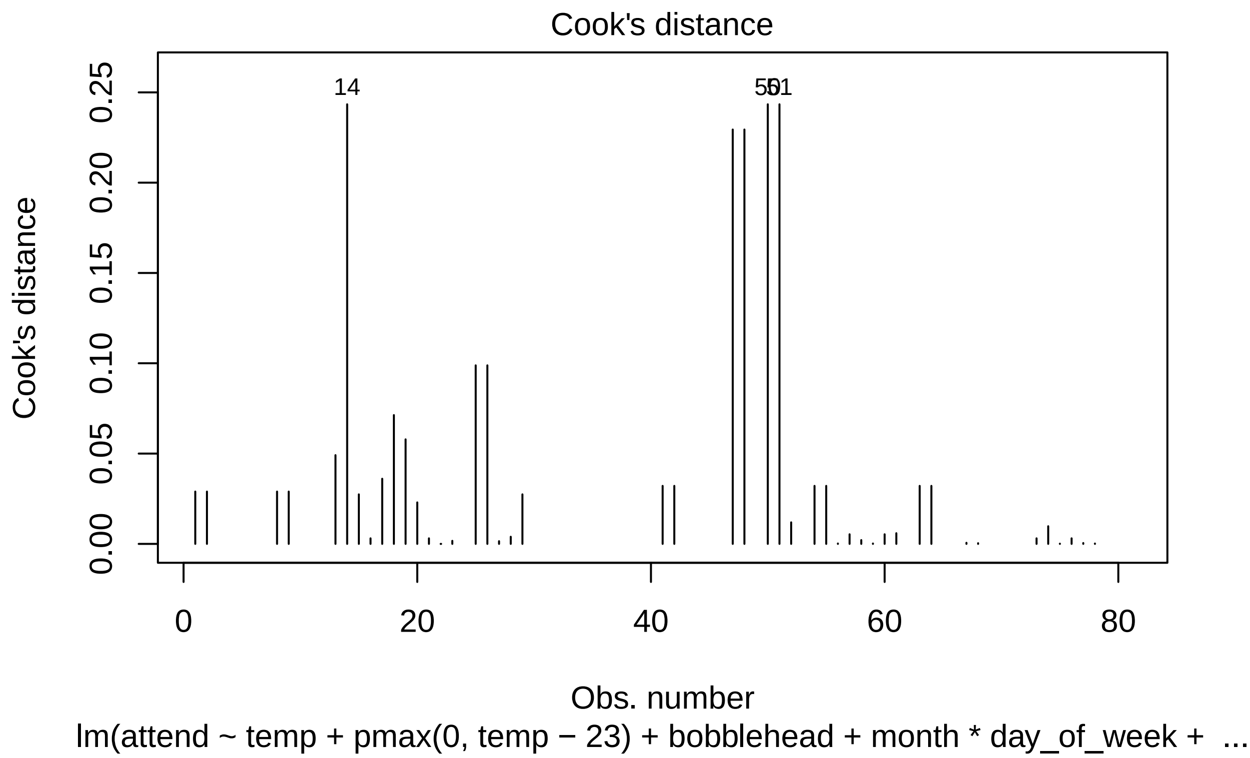
```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(model4)
## W = 0.95922, p-value = 0.1141
```

Null hypothesis says residuals have normal distribution and because p-value > 0.05, the test supports our inference and says the residuals are normally distributed. So, **normality assumption is met**.

The last condition is **constant variability**, which says that variability of points around the least squares line should be roughly constant. This implies that the variability of residuals around the zero line should be roughly constant as well.

This condition is also called *homoscedasticity* and we can check this using a residuals plot.

```
plot(model4, which=4)
```



This plot shows the relationship between the square root of the standardized residuals and the fitted values. The residuals are standardized by dividing by the MSE of the fitted line. The square root is taken to stabilize the variance of the residuals. From this graph, we check if there is a constant variance along residuals. Constant line means constant variance. Ideally, the points should be randomly distributed around a horizontal line with no discernible pattern. If there is a pattern in the residuals, this could indicate heteroscedasticity. Heteroscedasticity is a violation of the assumption of equal variance in a linear regression model. In our graph, there is not any strong trend but there are some fluctuations. So, it is hard to visually reach a conclusion. To be sure, it is useful to benefit from Non-Constant Variance Test.

```
ncvTest(model4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.077977, Df = 1, p = 0.079359
```

Null hypothesis says the variance is constant. Since p is big (>.05), the test says that the variance is constant along the residuals.

Also, check if there is any **influential point**.

```
# Generate Cook's distance values
cooksd <- cooks.distance(model4)

# Identify influential observations
which(cooksd > 1)
```

```
## named integer(0)
```

So, there is **no influential point**.

Diagnostic plots show that **the model meets the requirements of linear regression**.

Also, adjusted R2 values and AIC support that the best model is model4.

```
summary(model4)
```

```
##
## Call:
## lm(formula = attend ~ temp + pmax(0, temp - 23) + bobblehead +
##     month * day_of_week + temp * shirt + fireworks * opponent +
##     cap, data = events)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5746.5  -329.1     0.0   229.3  5806.5
##
## Coefficients: (20 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1704.3    29615.9   0.058  0.95498
## temp                              820.3     1135.0   0.723  0.48268
## pmax(0, temp - 23)               -805.5     1320.7  -0.610  0.55245
## bobbleheadYES                   17700.5     5816.5   3.043  0.00942 **
## monthMAY                        16760.0    21332.6   0.786  0.44616
## monthJUN                         9587.2    15010.5   0.639  0.53411
## monthJUL                        12701.0    22903.9   0.555  0.58863
## monthAUG                        11544.0    21378.6   0.540  0.59834
## monthSEP                         2724.7    14982.1   0.182  0.85849
## monthOCT                         8474.9    23720.7   0.357  0.72662
## day_of_weekTuesday              17737.9     7008.4   2.531  0.02508 *
## day_of_weekWednesday            -2591.6     6433.4  -0.403  0.69361
## day_of_weekThursday             -4912.7     8466.9  -0.580  0.57168
## day_of_weekFriday              -10787.2    15897.6  -0.679  0.50933
## day_of_weekSaturday             23204.9    12308.7   1.885  0.08194 .
## day_of_weekSunday               11733.8     9761.3   1.202  0.25078
## shirtYES                        22862.3    12383.6   1.846  0.08776 .
## fireworksYES                    20684.5    10398.3   1.989  0.06814 .
## opponentAstros                  -4456.5    27411.0  -0.163  0.87335
## opponentBraves                  11547.5    21472.3   0.538  0.59981
## opponentBrewers                   127.8    26868.7   0.005  0.99628
## opponentCardinals                 728.2    26862.1   0.027  0.97879
## opponentCubs                     5790.4    31989.4   0.181  0.85915
## opponentGiants                   4475.2    27174.4   0.165  0.87172
## opponentMarlins                  -957.9    31540.4  -0.030  0.97623
## opponentMets                     8114.5    17316.2   0.469  0.64711
## opponentNationals               -6413.4    19568.9  -0.328  0.74833
## opponentPadres                  10156.2    20827.2   0.488  0.63392
## opponentPhillies                 7189.7    29662.3   0.242  0.81226
## opponentPirates                 20052.8    22196.2   0.903  0.38273
## opponentReds                     1819.9    28650.7   0.064  0.95032
## opponentRockies                   729.7    26807.6   0.027  0.97870
## opponentSnakes                  -3990.7    27284.3  -0.146  0.88596
## opponentWhite Sox               11612.7    16459.1   0.706  0.49292
## capYES                          -8259.9     6894.6  -1.198  0.25230
## monthMAY:day_of_weekTuesday    -22859.6     8402.7  -2.721  0.01750 *
## monthJUN:day_of_weekTuesday     -7035.8    15371.9  -0.458  0.65472
```

```
## monthJUL:day_of_weekTuesday     -8002.2     8192.0  -0.977  0.34648
## monthAUG:day_of_weekTuesday    -14680.1     9741.0  -1.507  0.15571
## monthSEP:day_of_weekTuesday    -33476.7    14890.0  -2.248  0.04254 *
## monthOCT:day_of_weekTuesday     -8859.3     9886.5  -0.896  0.38649
## monthMAY:day_of_weekWednesday   -6058.8     7869.0  -0.770  0.45509
## monthJUN:day_of_weekWednesday   19209.2    16213.5   1.185  0.25732
## monthJUL:day_of_weekWednesday    2026.5    10279.3   0.197  0.84676
## monthAUG:day_of_weekWednesday    7974.3     7867.1   1.014  0.32926
## monthSEP:day_of_weekWednesday   19670.8     9764.8   2.014  0.06513 .
## monthOCT:day_of_weekWednesday    3011.2     9490.6   0.317  0.75607
## monthMAY:day_of_weekThursday    -4131.8    10391.0  -0.398  0.69735
## monthJUN:day_of_weekThursday    -2068.6    16701.8  -0.124  0.90333
## monthJUL:day_of_weekThursday         NA         NA      NA       NA
## monthAUG:day_of_weekThursday    13650.5    12142.0   1.124  0.28124
## monthSEP:day_of_weekThursday    24139.4    14077.3   1.715  0.11011
## monthOCT:day_of_weekThursday         NA         NA      NA       NA
## monthMAY:day_of_weekFriday      -7680.0    14640.8  -0.525  0.60872
## monthJUN:day_of_weekFriday      -7954.5     9950.8  -0.799  0.43843
## monthJUL:day_of_weekFriday      -9466.6    22658.6  -0.418  0.68292
## monthAUG:day_of_weekFriday       -428.5    11456.7  -0.037  0.97073
## monthSEP:day_of_weekFriday       2596.3    12108.4   0.214  0.83354
## monthOCT:day_of_weekFriday           NA         NA      NA       NA
## monthMAY:day_of_weekSaturday   -18601.2    13687.7  -1.359  0.19727
## monthJUN:day_of_weekSaturday   -17304.3     9754.7  -1.774  0.09948 .
## monthJUL:day_of_weekSaturday   -30333.6    23167.1  -1.309  0.21309
## monthAUG:day_of_weekSaturday   -13479.1     9761.4  -1.381  0.19060
## monthSEP:day_of_weekSaturday    -6042.4     9223.4  -0.655  0.52382
## monthOCT:day_of_weekSaturday         NA         NA      NA       NA
## monthMAY:day_of_weekSunday      -4835.7    11697.4  -0.413  0.68605
## monthJUN:day_of_weekSunday           NA         NA      NA       NA
## monthJUL:day_of_weekSunday     -15475.9    21263.5  -0.728  0.47963
## monthAUG:day_of_weekSunday           NA         NA      NA       NA
## monthSEP:day_of_weekSunday           NA         NA      NA       NA
## monthOCT:day_of_weekSunday           NA         NA      NA       NA
## temp:shirtYES                        NA         NA      NA       NA
## fireworksYES:opponentAstros      7753.9     8072.8   0.960  0.35433
## fireworksYES:opponentBraves          NA         NA      NA       NA
## fireworksYES:opponentBrewers         NA         NA      NA       NA
## fireworksYES:opponentCardinals   4146.4     5970.3   0.695  0.49959
## fireworksYES:opponentCubs       -3836.5     9075.9  -0.423  0.67941
## fireworksYES:opponentGiants          NA         NA      NA       NA
## fireworksYES:opponentMarlins         NA         NA      NA       NA
## fireworksYES:opponentMets       10368.4     9891.3   1.048  0.31363
## fireworksYES:opponentNationals  24033.8    11561.7   2.079  0.05801 .
## fireworksYES:opponentPadres          NA         NA      NA       NA
## fireworksYES:opponentPhillies        NA         NA      NA       NA
## fireworksYES:opponentPirates         NA         NA      NA       NA
## fireworksYES:opponentReds            NA         NA      NA       NA
## fireworksYES:opponentRockies         NA         NA      NA       NA
## fireworksYES:opponentSnakes          NA         NA      NA       NA
## fireworksYES:opponentWhite Sox       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4844 on 13 degrees of freedom
## Multiple R-squared:  0.9446, Adjusted R-squared:  0.6592
## F-statistic: 3.309 on 67 and 13 DF,  p-value: 0.0102
```

Question - 1

Now we will answer the question "does bobblehead still increase the attendance?".

We can answer this by checking the coefficient of bobblehead in the final model.

Expected number of additional fans drawn to a home game with a bobblehead promotion;

```
model4 %>% coef %>% .["bobbleheadYES"]
```

```
## bobbleheadYES
##       17700.5
```

So, the point estimate is 17701.

The confidence interval is (95%);

```
confint(model4, parm= "bobbleheadYES")
```

```
##                  2.5 %   97.5 %
## bobbleheadYES 5134.746 30266.26
```

Question - 2

Using our model, we can predict the number of attendees to a typical home game, - on a Wednesday,

- in June,
- the bobblehead promotion is applied,
- opponent is Angels,
- day_night is day,
- skies is clear,
- day is 4,
- temperature is 24,
- other promotions are not applied.

Besides point estimate, give a 90% prediction interval.

```
prediction_data <- data.table(day_of_week = "Wednesday",
                       month = "JUN",
                       bobblehead = "YES",
                       opponent = "Angels",
                       day_night = "Day",
                       skies = "Clear",
                       day = "4",
                       temp = 24,
                       shirt = "NO",
                       fireworks = "NO",
                       cap = "NO")

predict(object=model4, prediction_data)
```

```
## Warning in predict.lm(object = model4, prediction_data): prediction from a
## rank-deficient fit may be misleading
```

```
##        1
## 64490.34
```

So, the point estimate is 64490. The 90% prediction interval is,

```r
predict(object=model4, prediction_data, interval = "prediction", level = 0.9)
```

```
##        fit      lwr      upr
## 1 64490.34 45032.33 83948.34
```