DSC630Assignment2BKudaimiRMD

Bilal Kudaimi

6/20/2021

The goal of this assignment is to find out through EDA and regression which days are the best for running a marketing campaign to increase game attendee number. We have a dataset of LA Dodgers games with information such as month, day, daily temperature, game opponent, weather, number of attendees, and whether items such as caps, shirts, fireworks, and bobbleheads are sold at the game.

We will run the marketing campaign on days with the highest attendee number to increase the campaign's audience. To find out which days have the highest attendee number, we will use regression to tell us which features of this dataset contribute the most to attendee number, and from this, we can tell what days (e.g. days in March, Saturdays, etc) have the most weight in the model, and thus, which days would be the best to run our campaign.

EDA will be conducted to gain insight into the data, then multiple regression will be conducted to find out which features of the dataset weigh the most in predicting game attendee number.

Importing the data and viewing structure and summary statistics

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

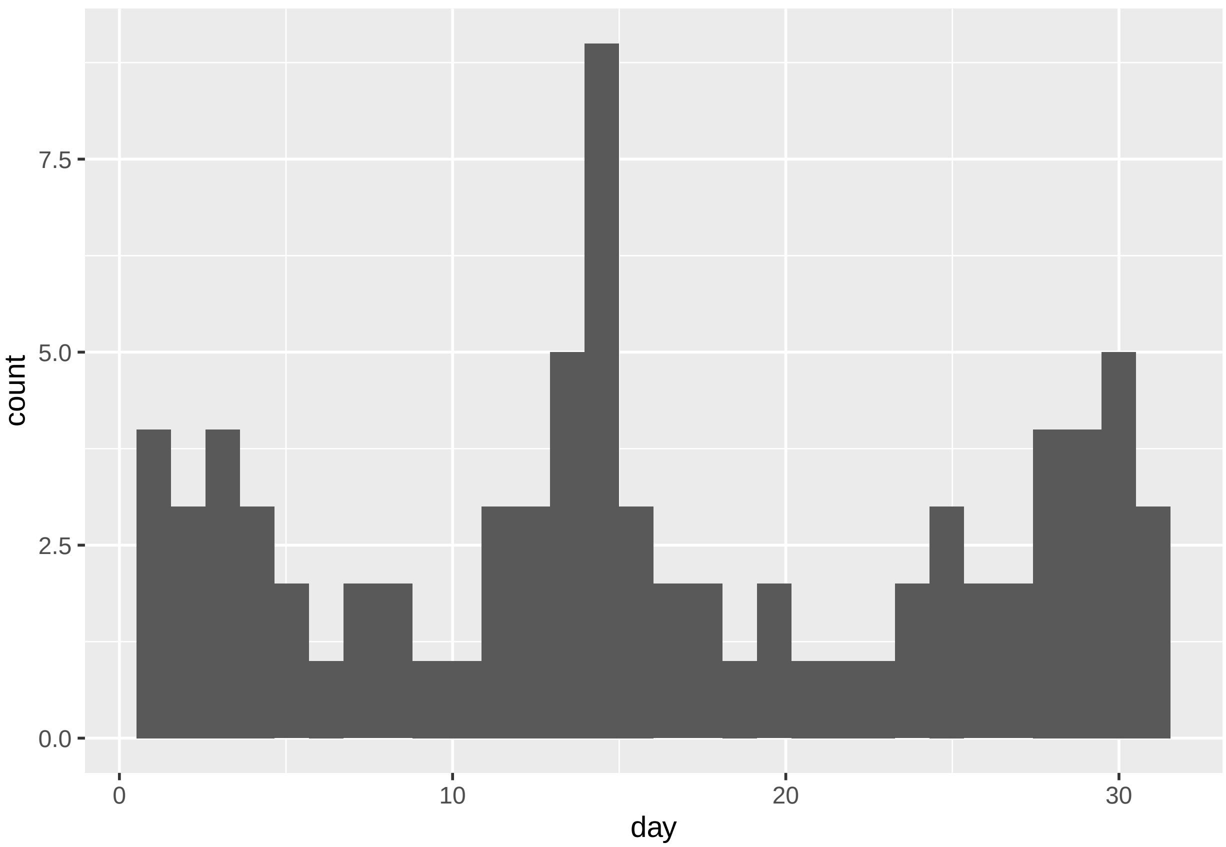
```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

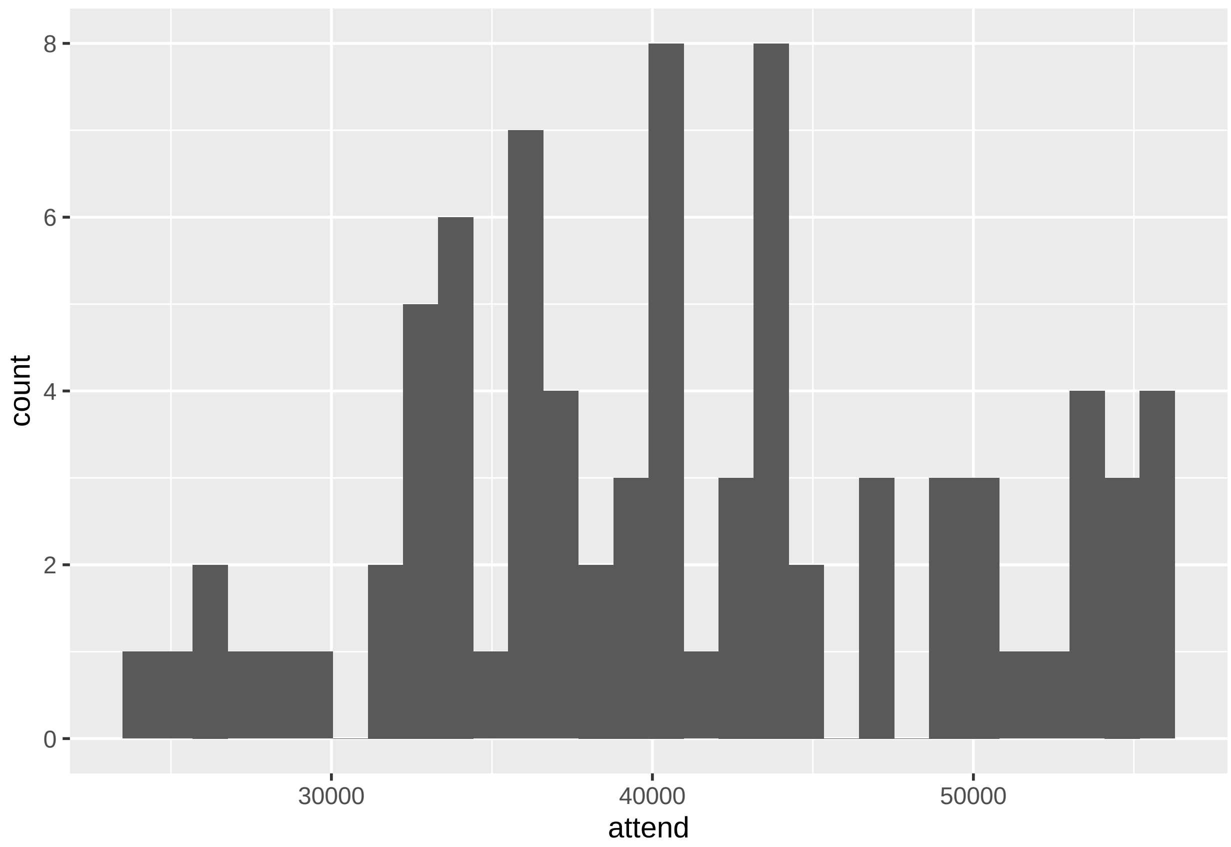
```
## Warning: package 'QuantPsyc' was built under R version 4.0.3
## Loading required package: boot

##
## Attaching package: ’boot’

## The following object is masked from ’package:survival’:
##
##     aml

## The following object is masked from ’package:lattice’:
##
##     melanoma

## Loading required package: MASS

##
## Attaching package: ’QuantPsyc’

## The following object is masked from ’package:base’:
##
##     norm

## [1] FALSE

## ’data.frame’:    81 obs. of  12 variables:
##  $ month      : chr  "APR" "APR" "APR" "APR" ...
##  $ day        : int  10 11 12 13 14 15 23 24 25 27 ...
##  $ attend     : int  56000 29729 28328 31601 46549 38359 26376 44014 26345 44807 ...
##  $ day_of_week: chr  "Tuesday" "Wednesday" "Thursday" "Friday" ...
##  $ opponent   : chr  "Pirates" "Pirates" "Pirates" "Padres" ...
##  $ temp       : int  67 58 57 54 57 65 60 63 64 66 ...
##  $ skies      : chr  "Clear " "Cloudy" "Cloudy" "Cloudy" ...
##  $ day_night  : chr  "Day" "Night" "Night" "Night" ...
##  $ cap        : chr  "NO" "NO" "NO" "NO" ...
##  $ shirt      : chr  "NO" "NO" "NO" "NO" ...
##  $ fireworks  : chr  "NO" "NO" "NO" "YES" ...
##  $ bobblehead : chr  "NO" "NO" "NO" "NO" ...

## dodgers
##
##  12  Variables      81  Observations
## --------------------------------------------------------------------------------
## month
##        n  missing distinct
##       81        0        7
##
## lowest : APR AUG JUL JUN MAY, highest: JUL JUN MAY OCT SEP
##
## Value        APR   AUG   JUL   JUN   MAY   OCT   SEP
## Frequency     12    15    12     9    18     3    12
## Proportion 0.148 0.185 0.148 0.111 0.222 0.037 0.148
## --------------------------------------------------------------------------------
## day
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       31    0.998    16.14     11.1        2        3
##      .25      .50      .75      .90      .95
##        8       15       25       29       30
##
## lowest :  1  2  3  4  5, highest: 27 28 29 30 31
## --------------------------------------------------------------------------------
## attend
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       80        1    41040     9525    26773    31607
##      .25      .50      .75      .90      .95
##    34493    40284    46588    53570    55024
##
## lowest : 24312 25509 26345 26376 26773, highest: 54621 55024 55279 55359 56000
## --------------------------------------------------------------------------------
## day_of_week
##        n  missing distinct
##       81        0        7
##
## lowest : Friday    Monday    Saturday  Sunday    Thursday
## highest: Saturday  Sunday    Thursday  Tuesday   Wednesday
##
## Value         Friday    Monday  Saturday    Sunday  Thursday   Tuesday
## Frequency         13        12        13        13         5        13
## Proportion     0.160     0.148     0.160     0.160     0.062     0.160
##
## Value      Wednesday
## Frequency         12
## Proportion     0.148
## --------------------------------------------------------------------------------
## opponent
##        n  missing distinct
##       81        0       17
##
## lowest : Angels    Astros    Braves    Brewers   Cardinals
## highest: Pirates   Reds      Rockies   Snakes    White Sox
##
## Angels (3, 0.037), Astros (3, 0.037), Braves (3, 0.037), Brewers (4, 0.049),
## Cardinals (7, 0.086), Cubs (3, 0.037), Giants (9, 0.111), Marlins (3, 0.037),
## Mets (4, 0.049), Nationals (3, 0.037), Padres (9, 0.111), Phillies (3, 0.037),
## Pirates (3, 0.037), Reds (3, 0.037), Rockies (9, 0.111), Snakes (9, 0.111),
## White Sox (3, 0.037)
## --------------------------------------------------------------------------------
## temp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       32    0.997    73.15    9.391       59       64
##      .25      .50      .75      .90      .95
##       67       73       79       84       86
##
## lowest : 54 57 58 59 60, highest: 84 85 86 89 95
## --------------------------------------------------------------------------------
## skies
##        n  missing distinct
##       81        0        2
##
## Value       Clear Cloudy
## Frequency      62     19
## Proportion  0.765  0.235
## --------------------------------------------------------------------------------
## day_night
##        n  missing distinct
##       81        0        2
##
## Value        Day Night
## Frequency     15    66
## Proportion 0.185 0.815
## --------------------------------------------------------------------------------
## cap
##        n  missing distinct
##       81        0        2
##
## Value         NO   YES
## Frequency     79     2
## Proportion 0.975 0.025
## --------------------------------------------------------------------------------
## shirt
##        n  missing distinct
##       81        0        2
##
## Value         NO   YES
## Frequency     78     3
## Proportion 0.963 0.037
## --------------------------------------------------------------------------------
## fireworks
##        n  missing distinct
##       81        0        2
##
## Value         NO   YES
## Frequency     67    14
## Proportion 0.827 0.173
## --------------------------------------------------------------------------------
## bobblehead
##        n  missing distinct
##       81        0        2
##
## Value         NO   YES
## Frequency     70    11
## Proportion 0.864 0.136
## --------------------------------------------------------------------------------
```

Generating histograms and bar charts of the continuous and categorical variables, respectively. Boxplots will also be generated for the numerical variables. This will reveal if there is any skewness among the variables.

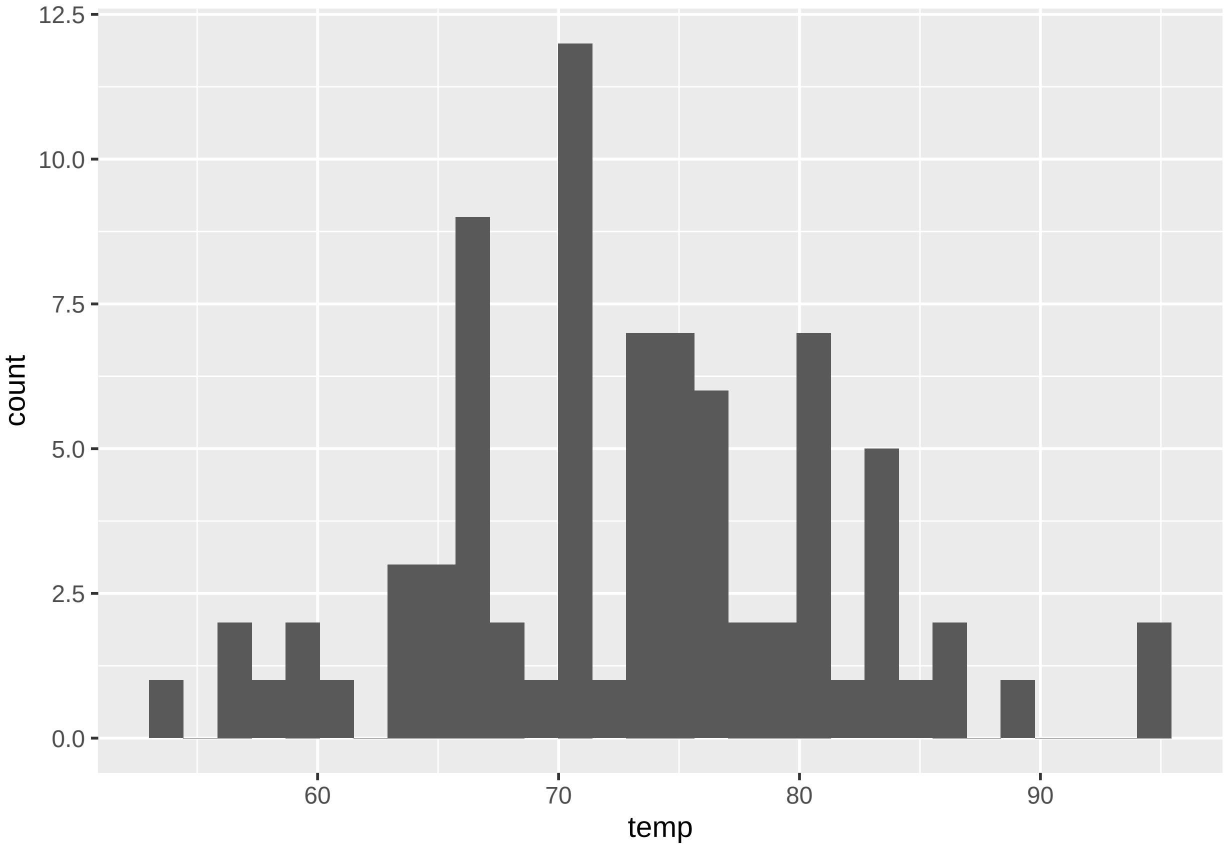
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

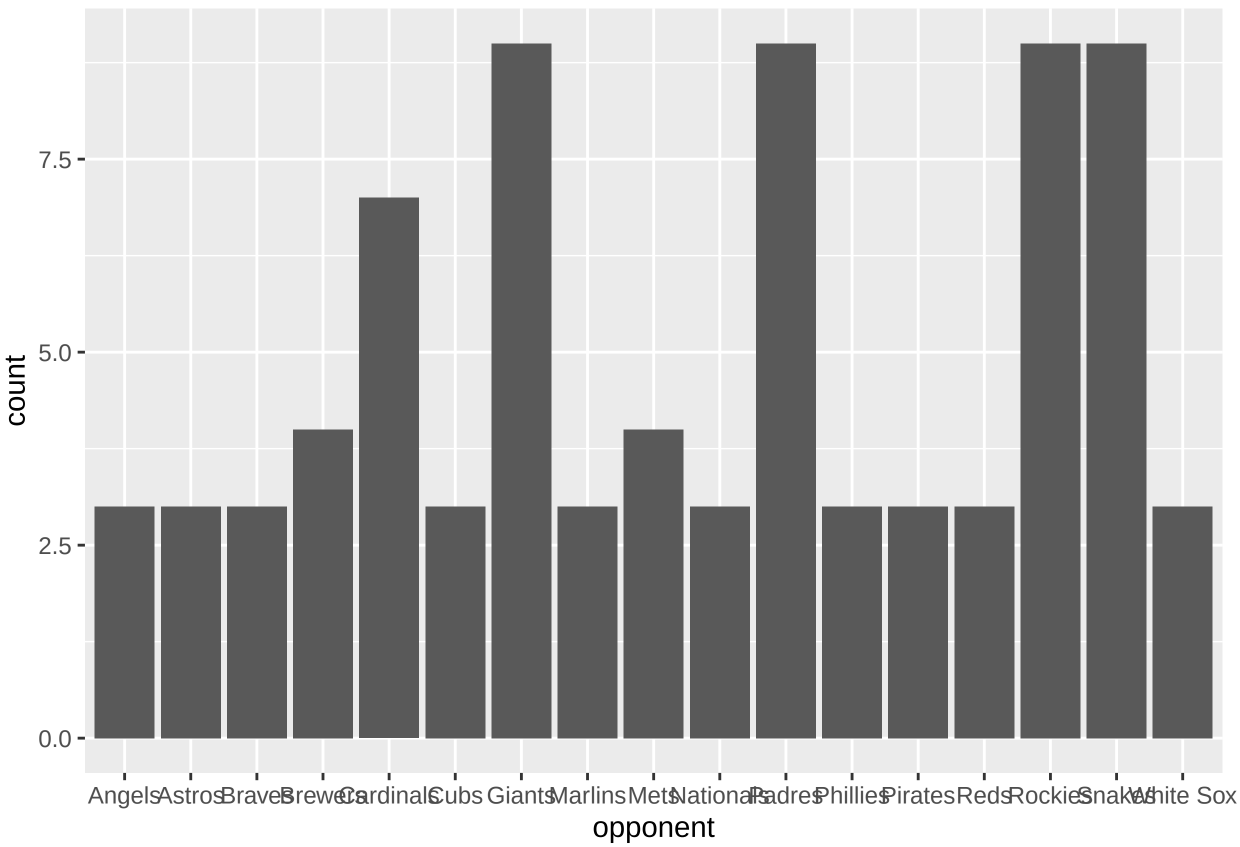


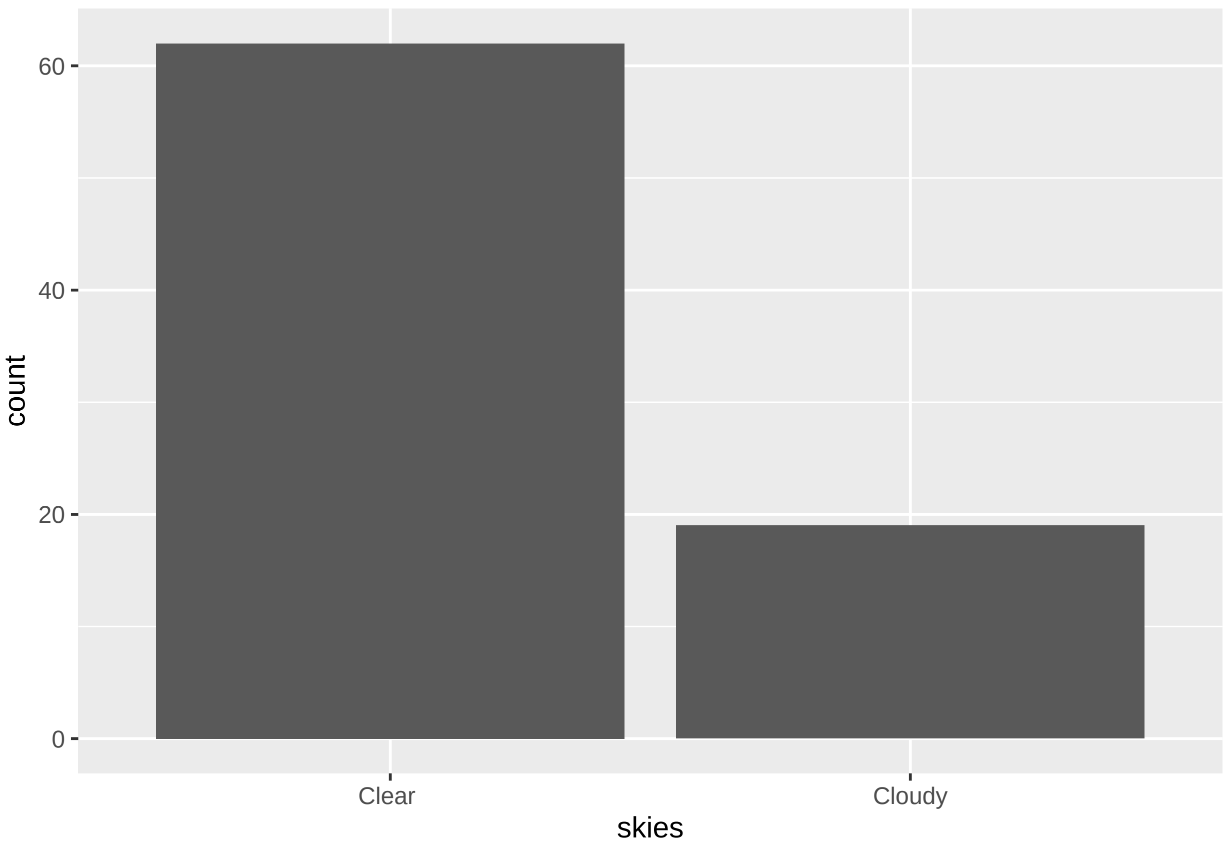
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

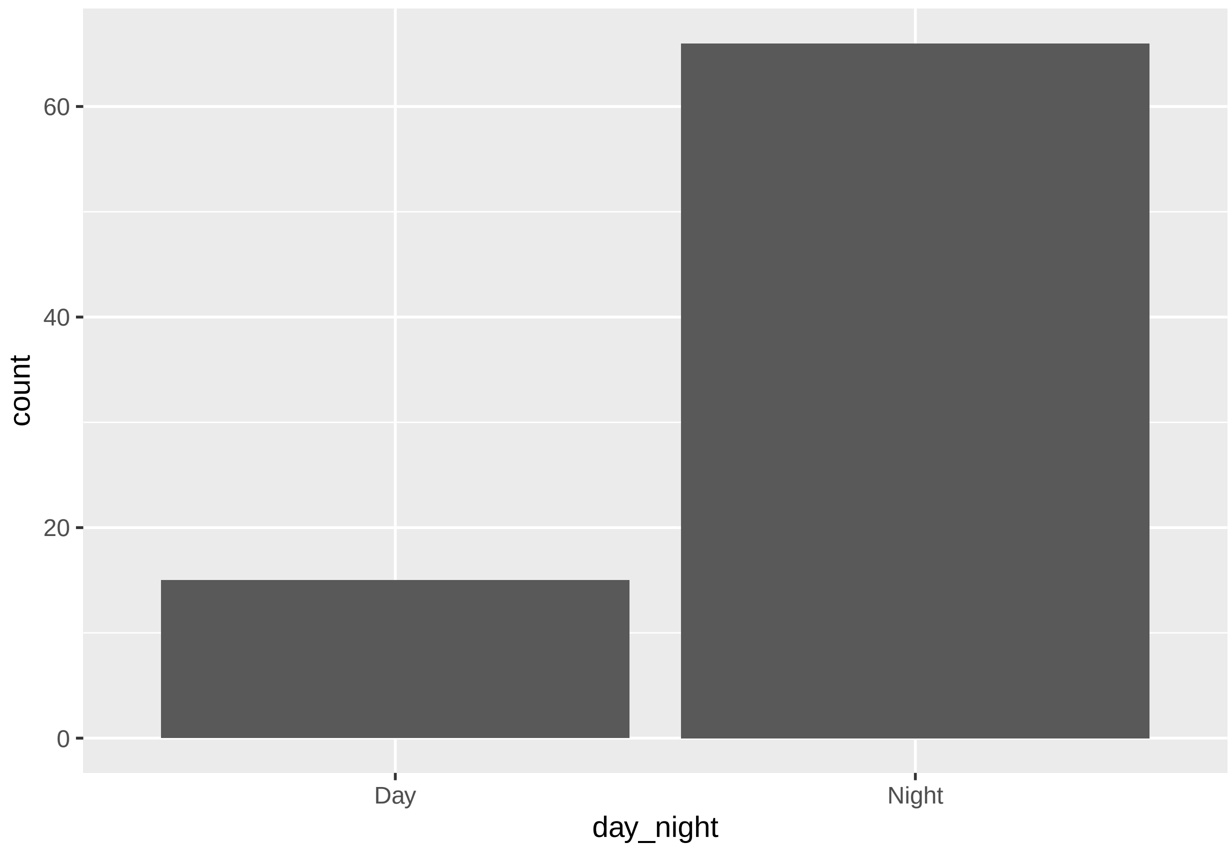


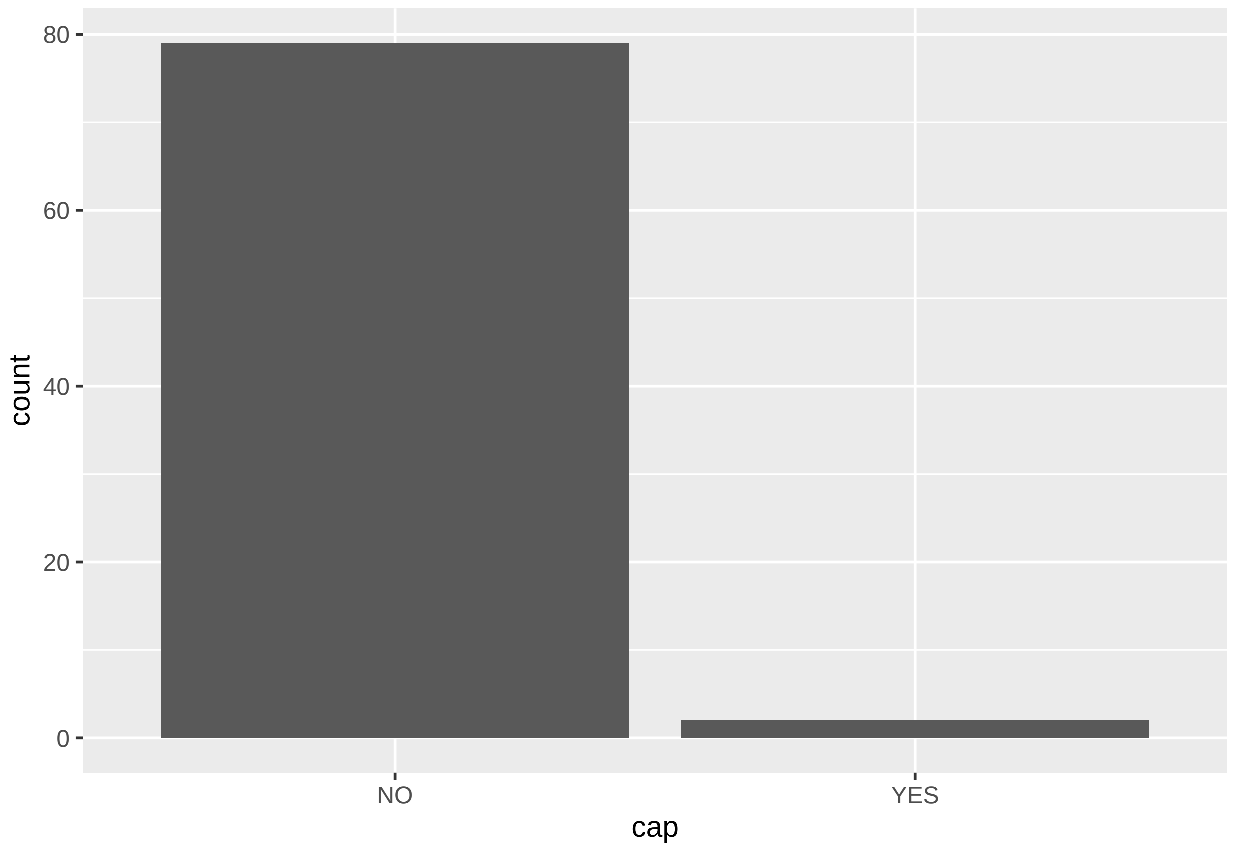
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

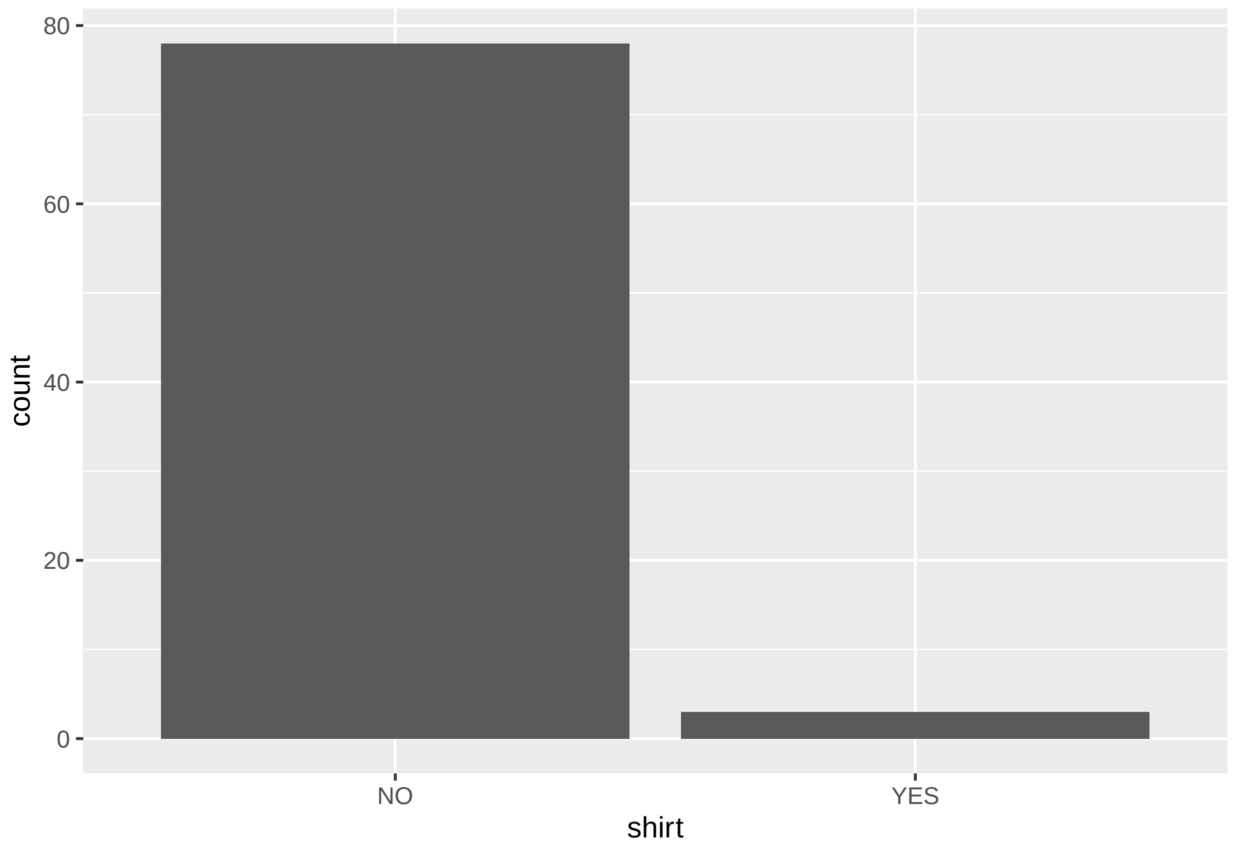

12.5
10.0
7.5
5.0
2.5
0.0
count
60
70
80
90
temp

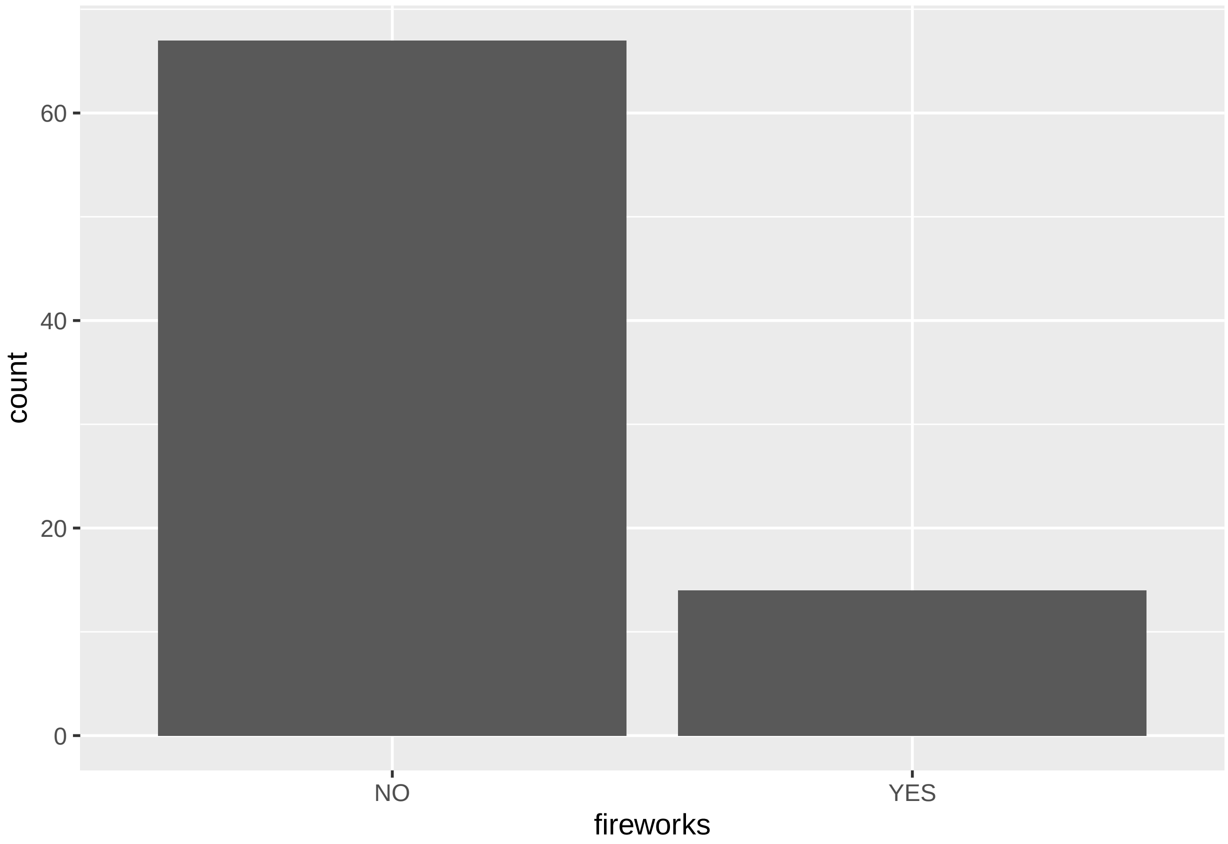

count
7.5
5.0
2.5
0.0
Angels
Astros
Braves
Brewers
Cardinals
Cubs
Giants
Marlins
Mets
Nationals
Padres
Phillies
Pirates
Reds
Rockies
Snakes
White Sox
opponent

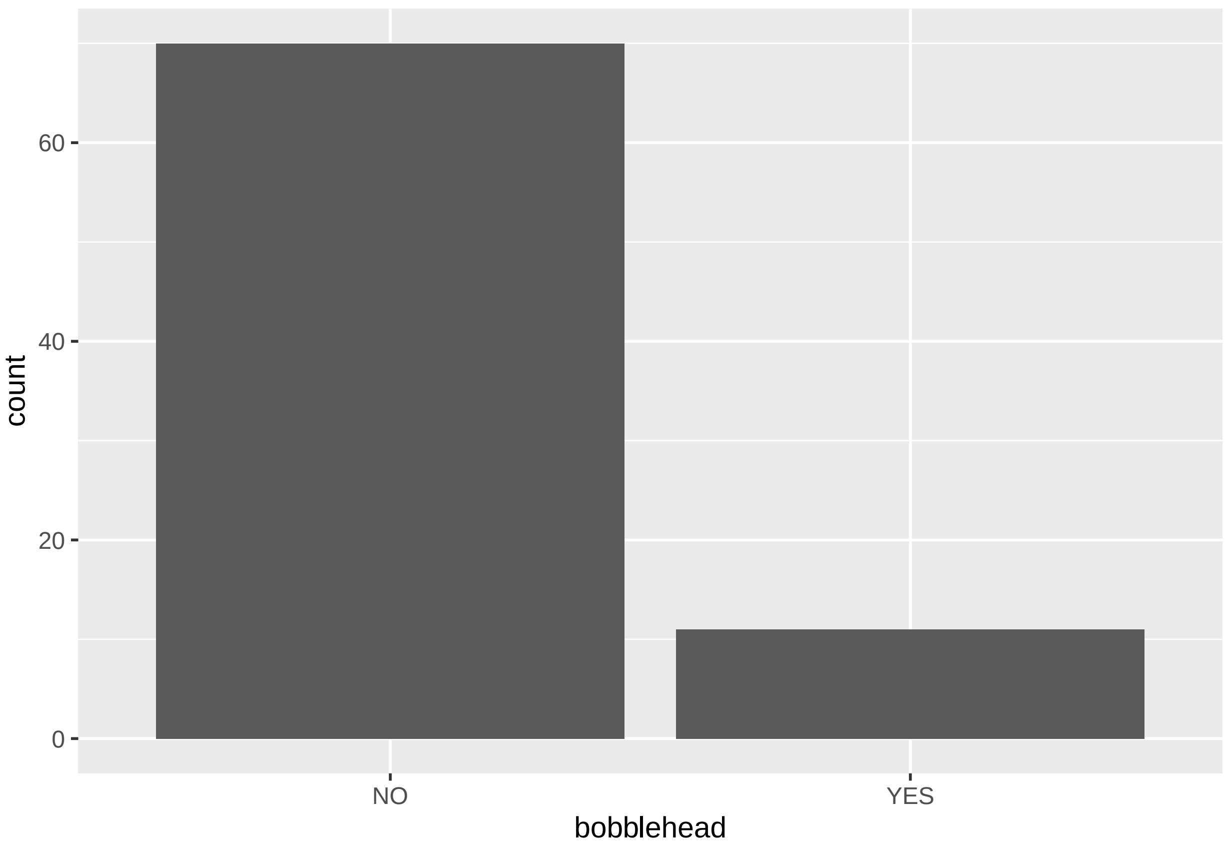

60
40
20
0
count
Clear
Cloudy
skies

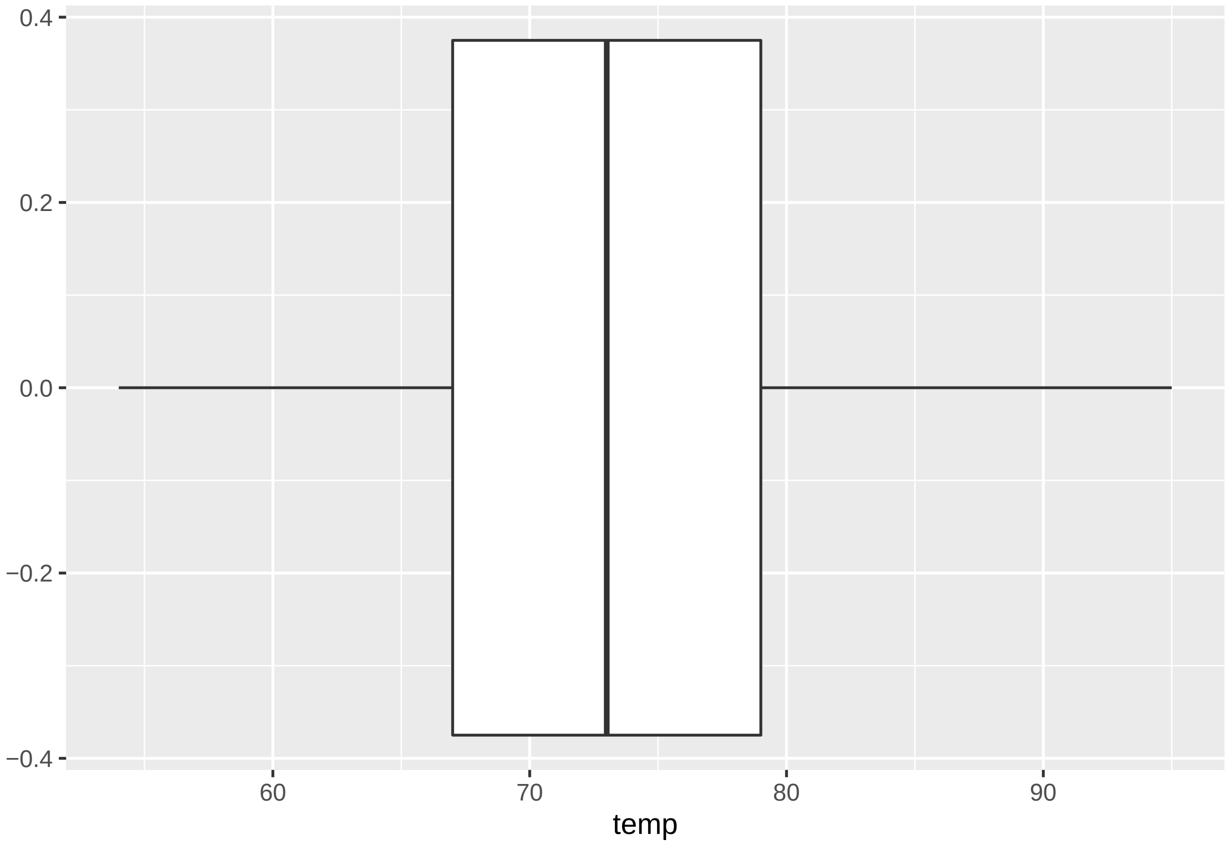

60
40
20
0
count
Day
Night
day_night

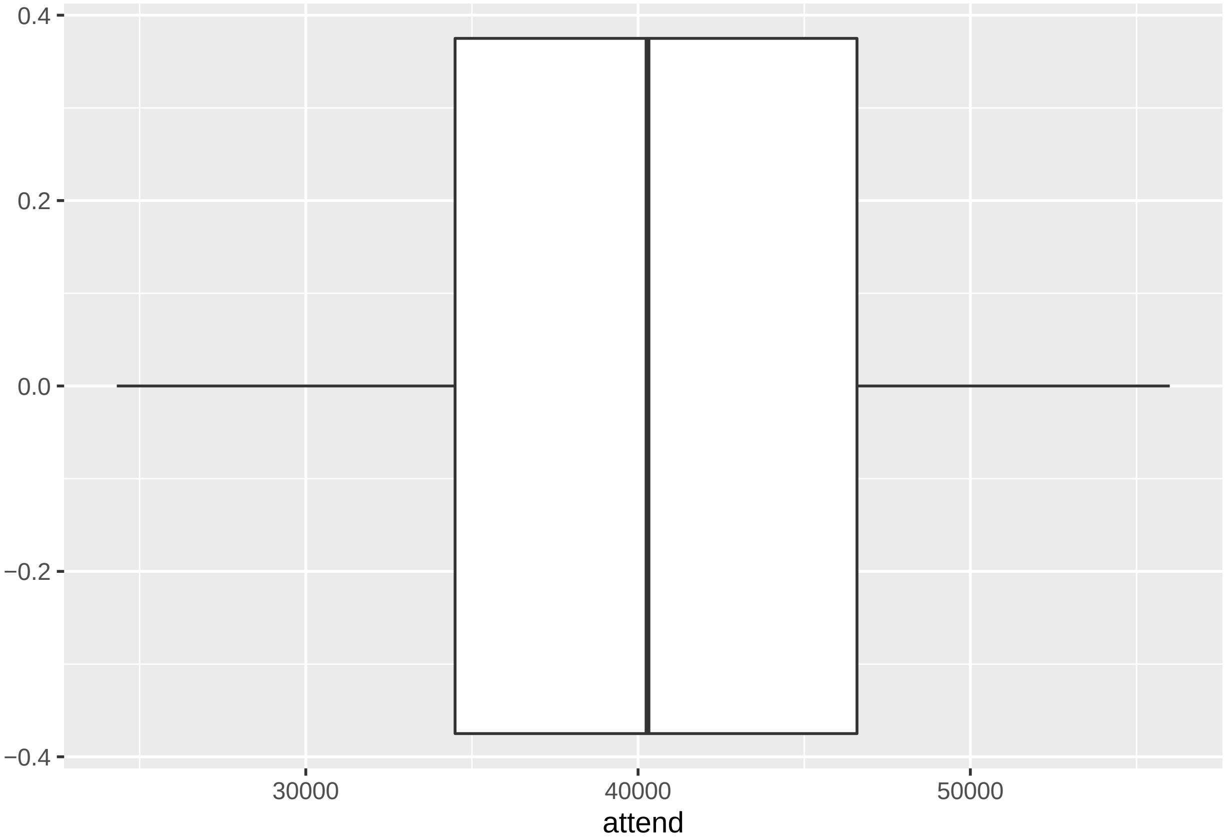

80
60
40
20
0
count
NO
YES
cap

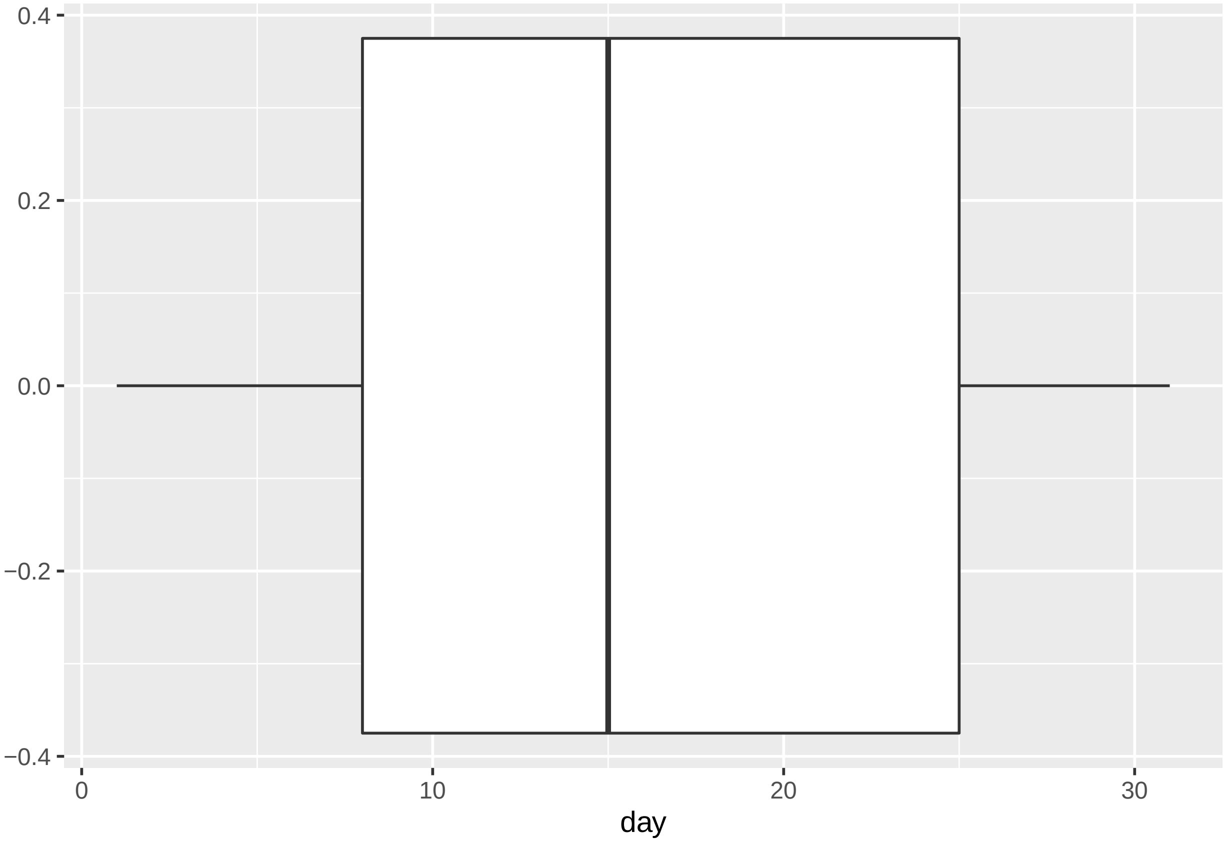

80
60
40
20
0
count
NO
YES
shirt


60
40
20
0
count
NO
YES
fireworks


60
40
20
0
count
NO
YES
bobblehead

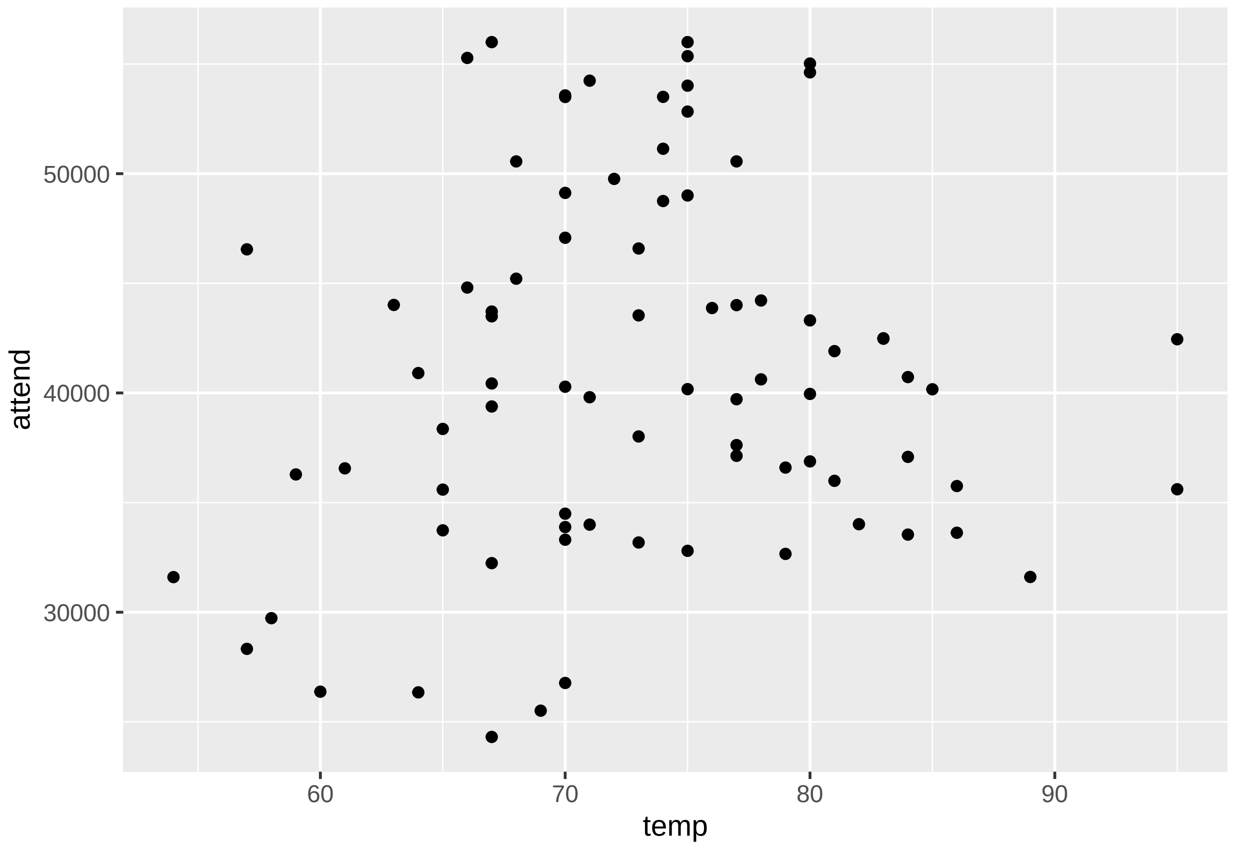

0.4
0.2
0.0
−0.2
−0.4
60
70
80
90
temp

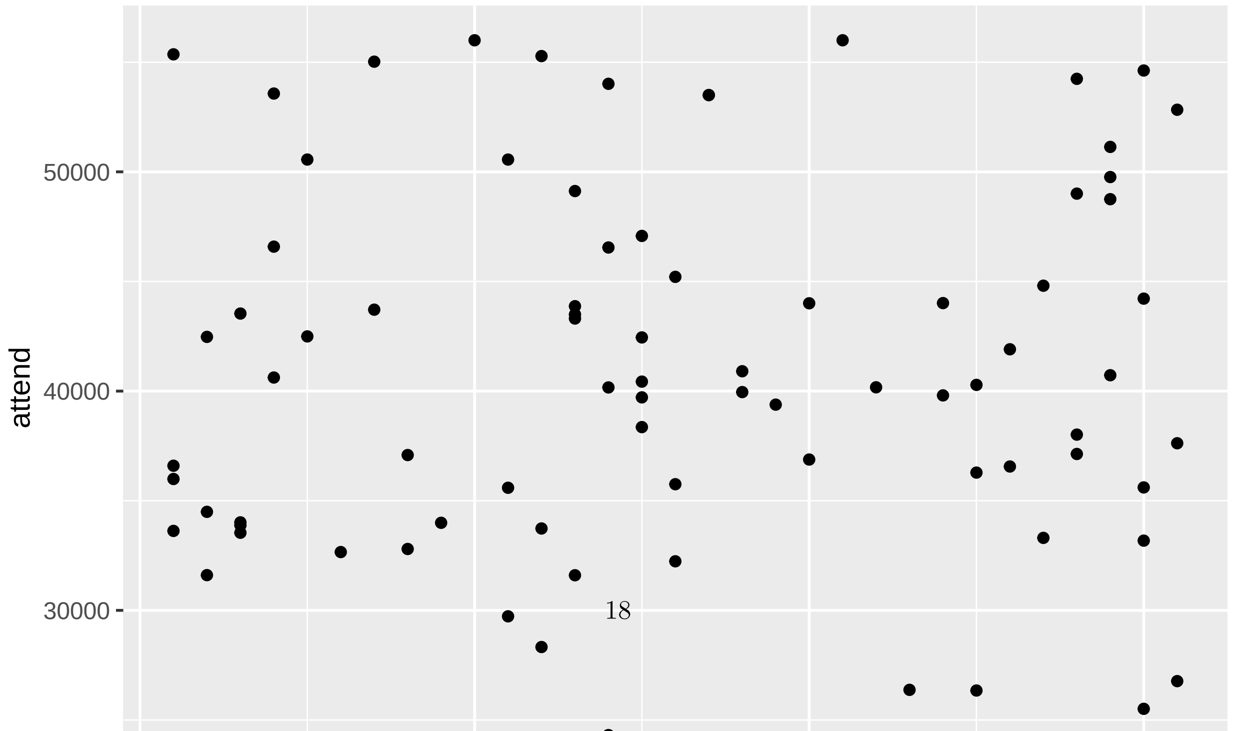

0.4
0.2
0.0
−0.2
−0.4
30000
40000
50000
attend


0.4
0.2
0.0
−0.2
−0.4
0
10
20
30
day

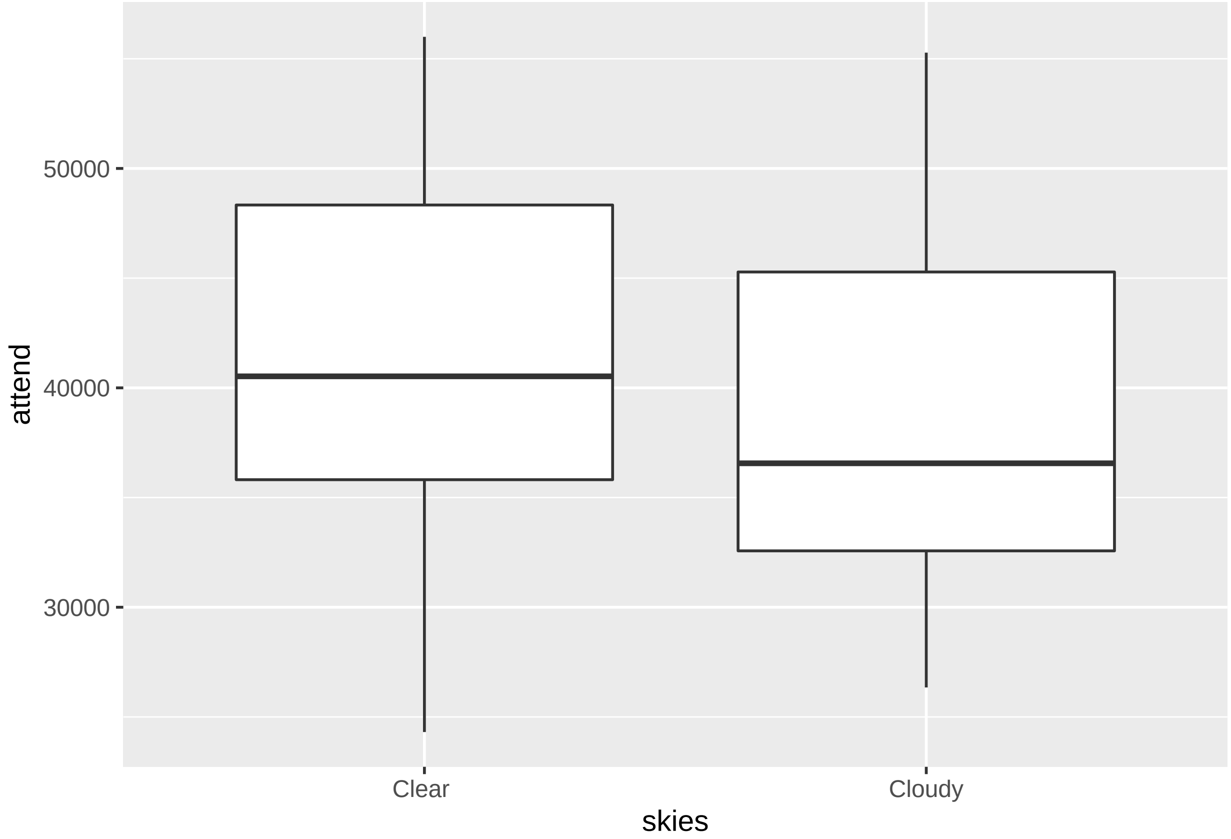
There does not appear to be any skewness, however, most of the games were played on clear nights and souvenirs such as caps, shirts, fireworks, and bobbleheads weren't sold at a majority of games. It also appears the 15th of each month is a popular day for attending baseball games, for some reason.

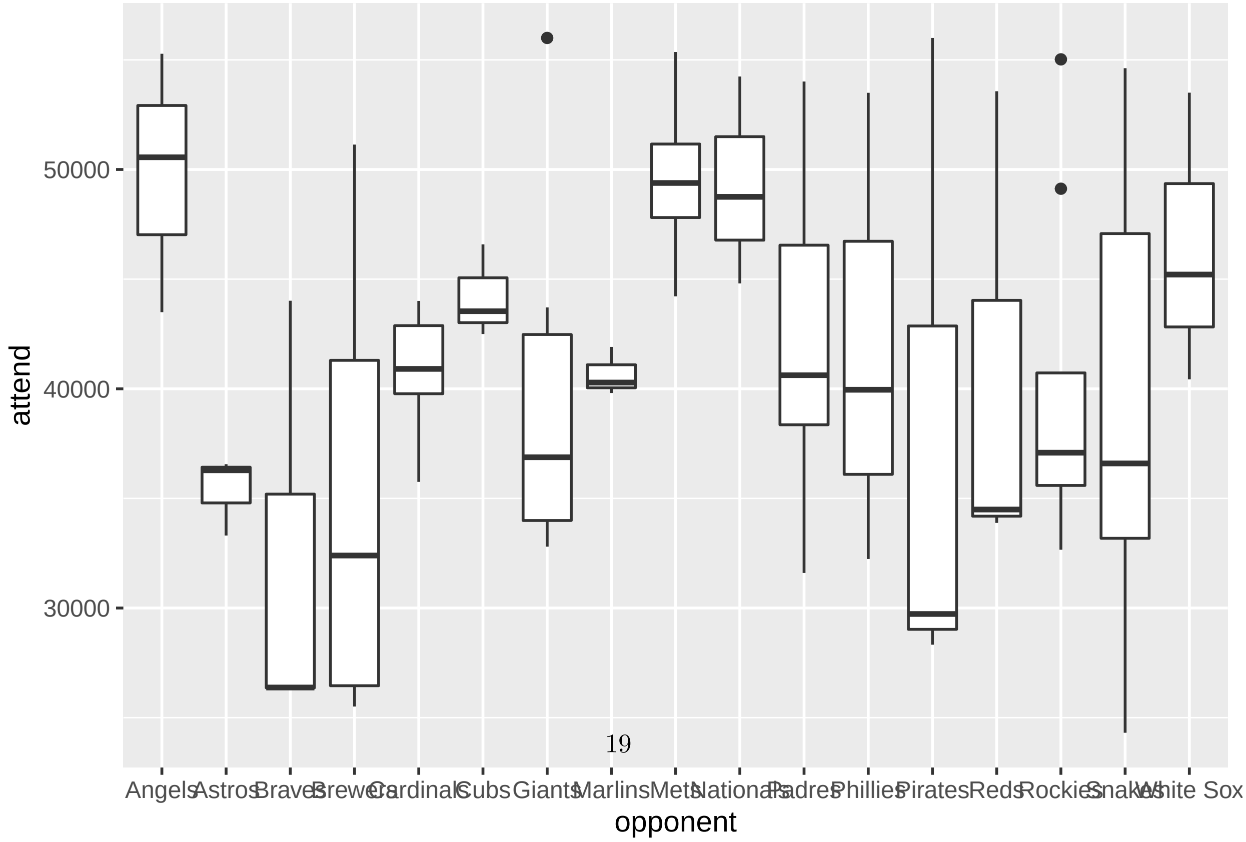
Two scatter plots will now be generated of game attendees vs temperature and day of the month to determine if temperature or day number affects the number of attendees.

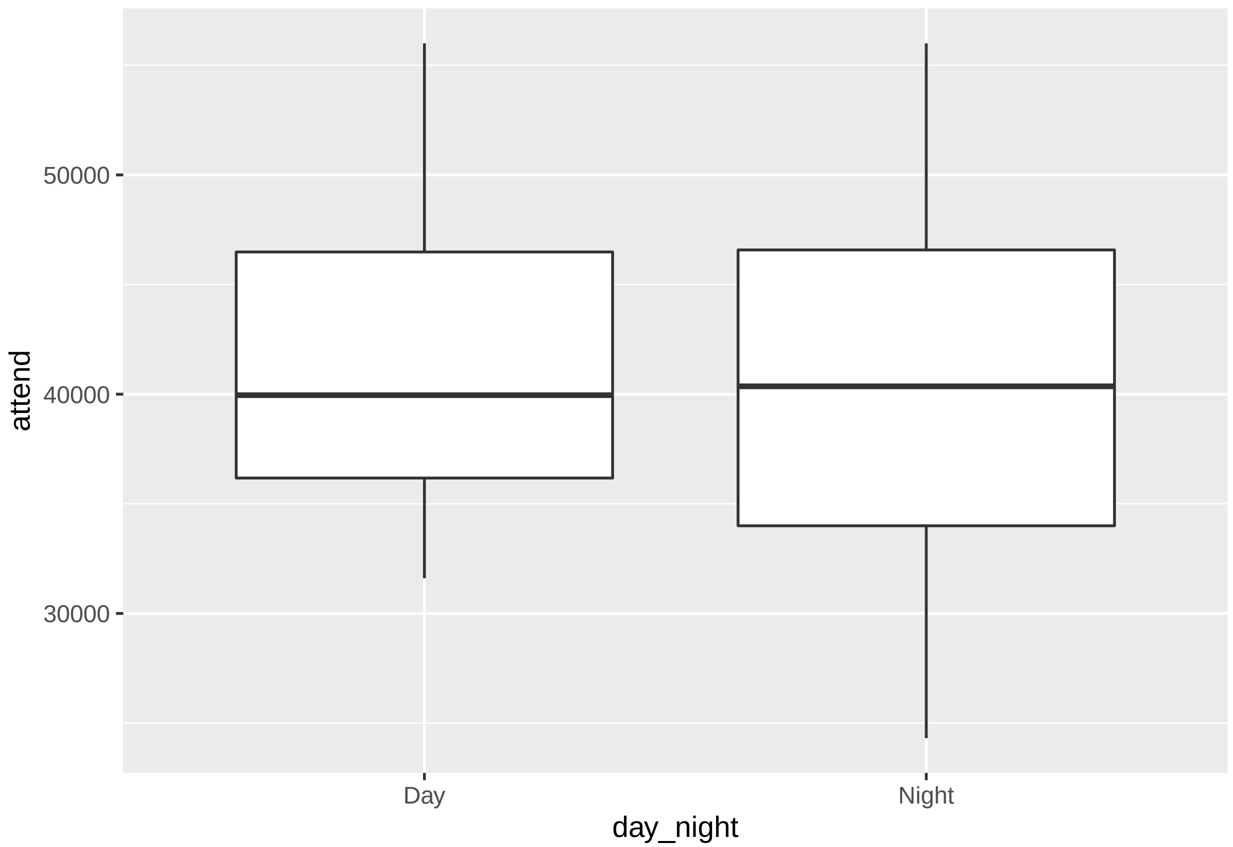


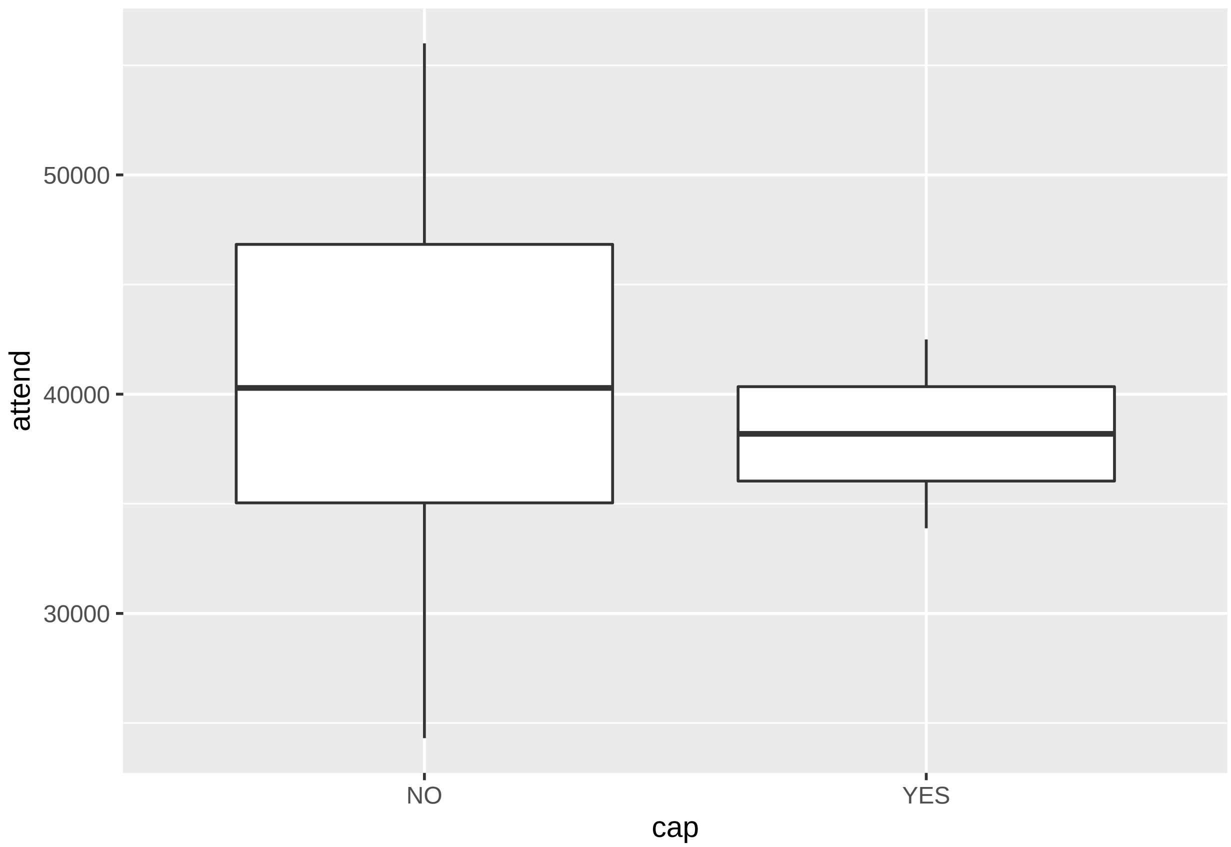


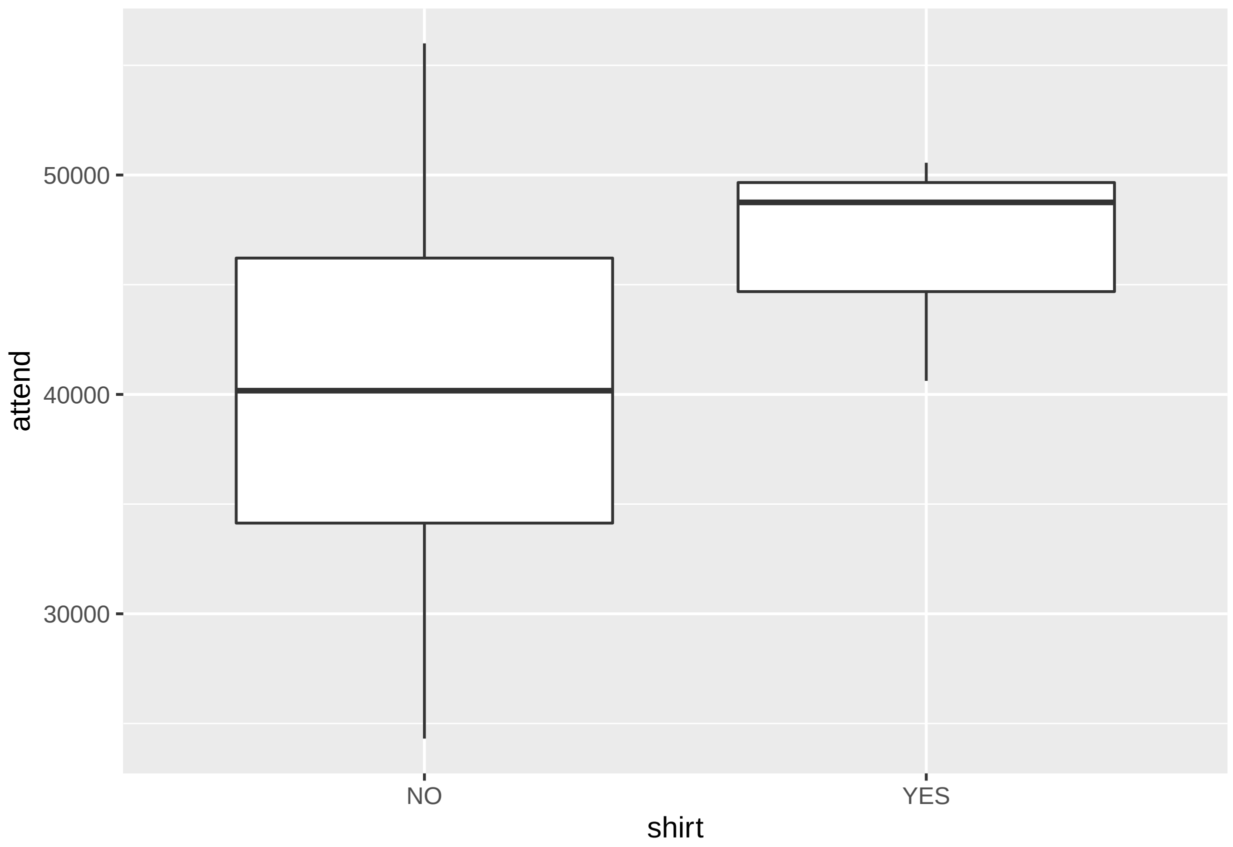
It doesn't look like temperature or day number is correlated with number of attendees, so the other variables will be plotted against attendee number to see if there are any correlations. Boxplots will be used since the remaining variables to be plotted are categorical in nature.

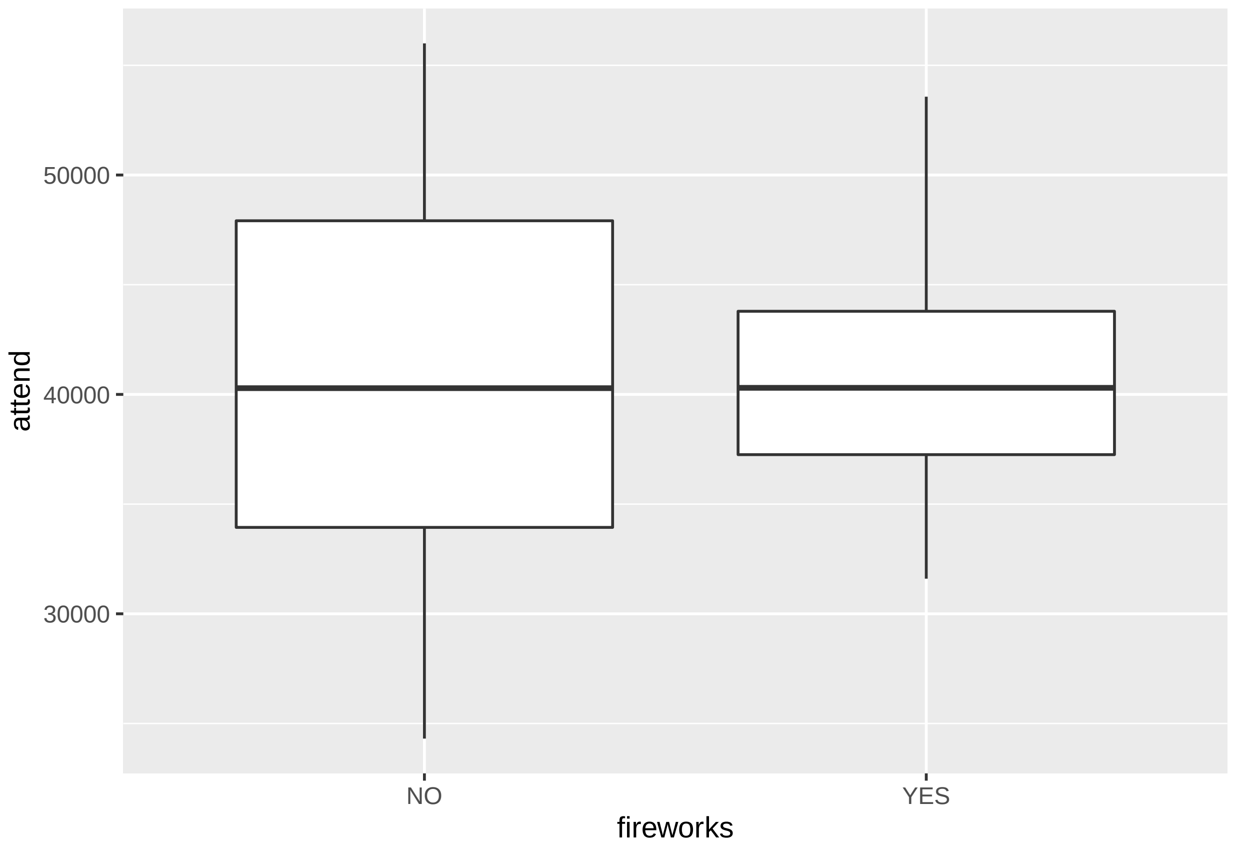


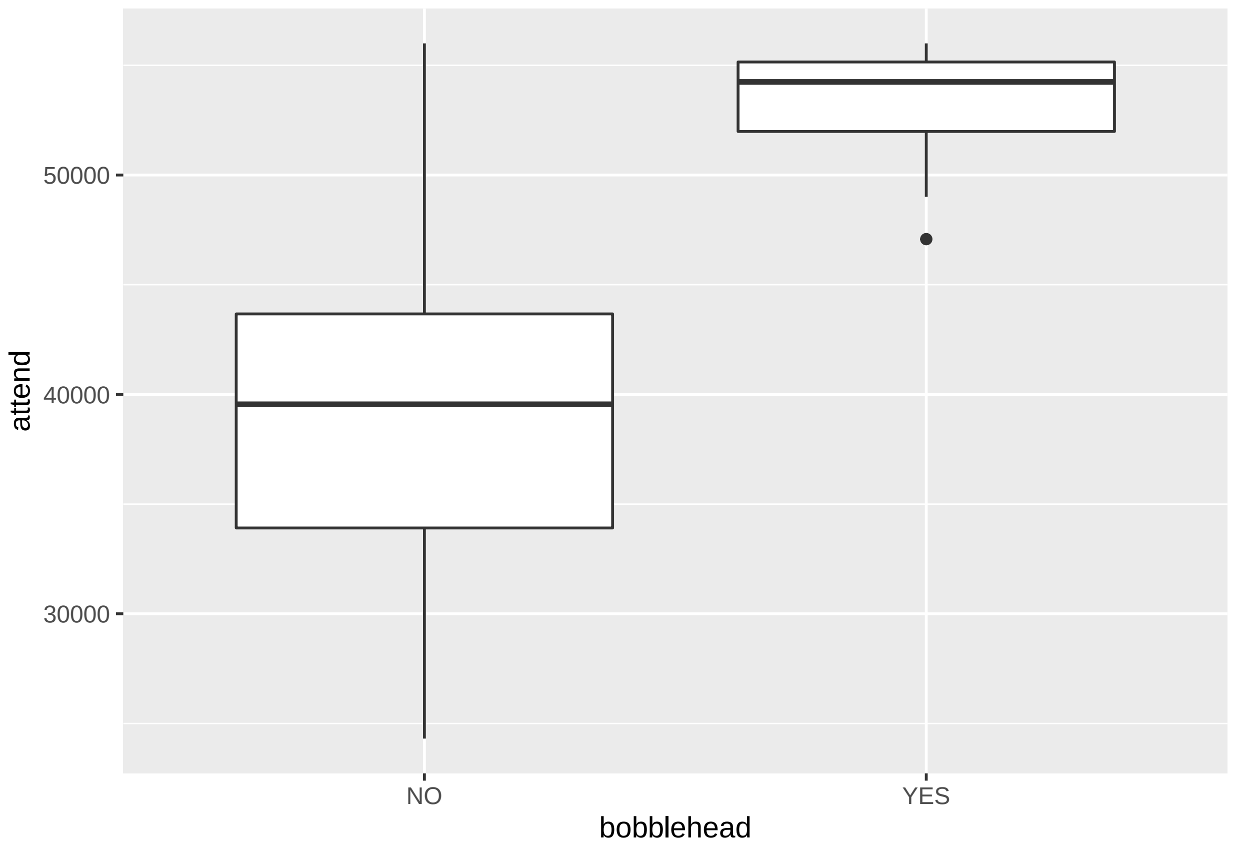


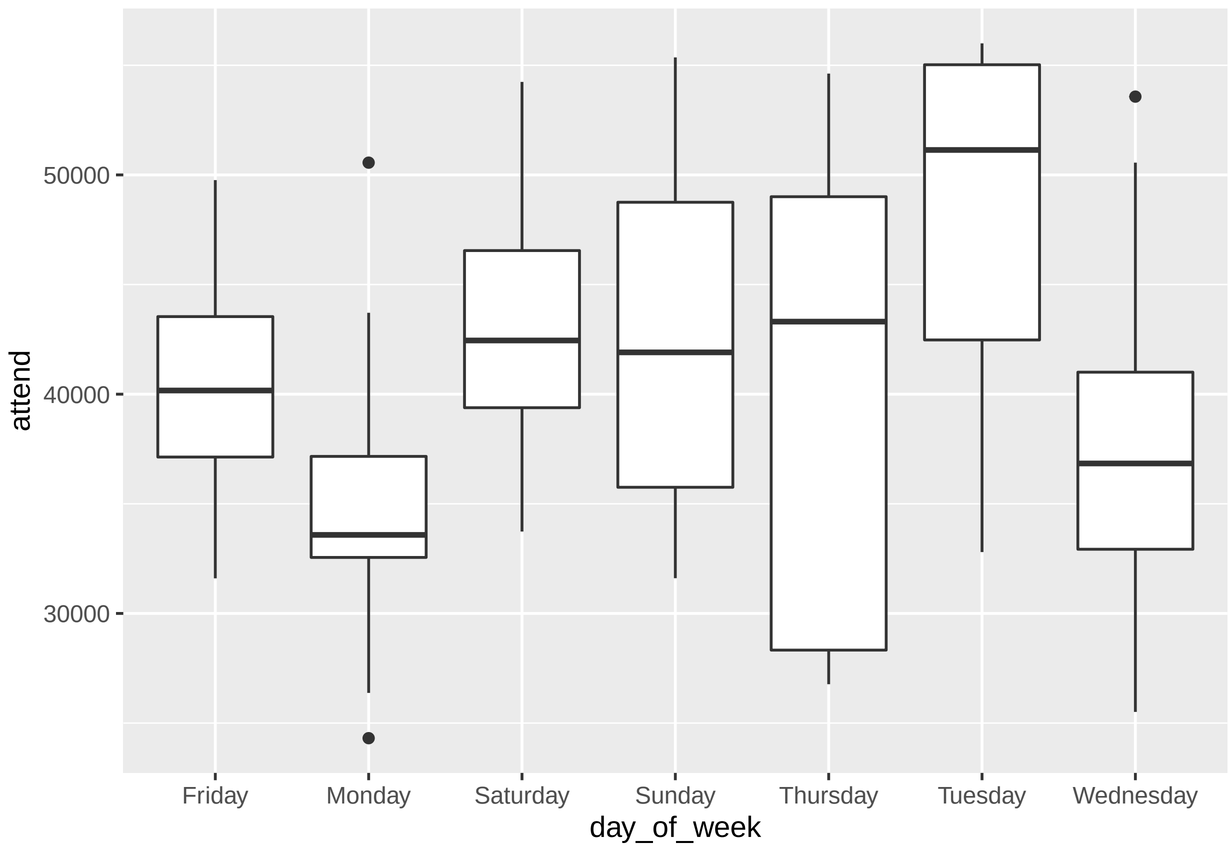

50000
40000
30000
attend
Day
Night
day_night


attend
50000
40000
30000
NO
YES
cap


50000
40000
30000
attend
NO
YES
shirt


50000
40000
30000
attend
NO
YES
fireworks


50000
40000
30000
attend
NO
YES
bobblehead



It appears that Tuesdays have the highest rate of attendance, and that bobbleheads and shirts present at a game are correlated with high attendance. It also appears that games vs the Angels, Mets, and Nationals garner the highest median attendee number, while games vs the Braves and the Pirates garner the lowest median attendee number. Based on the EDA, it appears that games on Tuesdays vs the Angels, Mets, or Nationals where shirts and bobbleheads are sold would be a good target for a marketing campaign.

Multiple linear regression will now be set up to determine which of the factors weigh into game attendee number the most. Any days of the week identified as having a significant weight in attendee number will be reported. First, though, it must be determined that none of the numeric variables are correlated with each other. Variables that have correlations with each other, or that have no correlation with the target variable (attendee number) will be dropped.

```
##                day     attend        temp
## day     1.00000000 0.02709298 -0.12761220
## attend  0.02709298 1.00000000  0.09895073
## temp   -0.12761220 0.09895073  1.00000000
```

None of the numeric features appear to be correlated with each other, so we can continue with building the model. Since temperature, day number, and skies did not appear to be correlated with attendee number, these variables will not be used in the regression model.

```
##
## Call:
## lm(formula = attend ~ month + day_of_week + opponent + cap +
##     shirt + fireworks + bobblehead, data = dodgers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9628.9 -2701.3    -1.1  1645.9 12822.6
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           20136.4    12229.2   1.647  0.10618
## monthAUG               7375.0     5794.5   1.273  0.20923
## monthJUL               4658.1     4684.5   0.994  0.32503
## monthJUN               5393.5     8676.0   0.622  0.53711
## monthMAY               2415.0     5358.2   0.451  0.65422
## monthOCT               2547.0     6766.3   0.376  0.70826
## monthSEP               2495.3     4806.7   0.519  0.60605
## day_of_weekMonday     17559.7     8830.6   1.989  0.05247 .
## day_of_weekSaturday   22152.6     8390.7   2.640  0.01115 *
## day_of_weekSunday     22503.1     8259.4   2.725  0.00896 **
## day_of_weekThursday   18307.6     8882.4   2.061  0.04473 *
## day_of_weekTuesday    26583.6     8935.6   2.975  0.00458 **
## day_of_weekWednesday  18038.8     8198.7   2.200  0.03264 *
## opponentAstros        -8801.2    10177.2  -0.865  0.39145
## opponentBraves        -8618.8    10260.2  -0.840  0.40506
## opponentBrewers       -9626.6     9859.4  -0.976  0.33376
## opponentCardinals     -2902.1     9662.6  -0.300  0.76521
## opponentCubs          -2824.2    10666.9  -0.265  0.79233
## opponentGiants        -6707.2     9612.6  -0.698  0.48870
## opponentMarlins       -8479.2    10507.4  -0.807  0.42366
## opponentMets          -1184.6     5407.4  -0.219  0.82752
## opponentNationals      3977.7     9748.8   0.408  0.68507
## opponentPadres        -2933.2     8778.6  -0.334  0.73974
## opponentPhillies      -3624.9     9457.1  -0.383  0.70319
## opponentPirates       -3094.1    10468.0  -0.296  0.76883
## opponentReds          -9507.1    10272.5  -0.925  0.35934
## opponentRockies       -6958.9     9453.1  -0.736  0.46522
## opponentSnakes        -9546.8     9068.0  -1.053  0.29770
## opponentWhite Sox      -781.1     5565.9  -0.140  0.88899
## capYES                -6341.2     5680.5  -1.116  0.26985
## shirtYES               1314.0     4420.0   0.297  0.76753
## fireworksYES          20243.7     8014.3   2.526  0.01489 *
## bobbleheadYES          9246.1     3030.6   3.051  0.00371 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5836 on 48 degrees of freedom
## Multiple R-squared:  0.7032, Adjusted R-squared:  0.5053
## F-statistic: 3.554 on 32 and 48 DF,  p-value: 3.7e-05
```

The model coefficients and significance codes will reveal how much weight each variable has in predicting attendee numbers. Based on these coefficients in the regression model summary, the days with the highest attendee number are Sundays and Tuesdays. In addition, any day when fireworks and bobbleheads are being sold have high attendee numbers. This matches the results of the EDA, although opponent did not weigh heavily into the regression model. In addition, the multiple R-squared value was 0.703, meaning that this model is a good fit for this data, and the p-value was 3.7e-5, meaning it is highly unlikely this model fit the data by chance. It is confirmed, then, that the best days to run our marketing campaign to spread it to a wider audience are Sundays, Tuesdays, and any day when fireworks and bobbleheads are sold.