Week 3 Exercise

Georges Essoh

9/18/2021

In the first part, we will focus on visualization. The idea is to identify the variables that could most influence the expectation. Indeed, by plotting the attendance against the other variables, one might detect a pattern between them and those that directly influence attendance. Second, we will use a linear regression model to confirm that the variables identified in the first part have had a significant impact on expectation.

1)Load Packages

```
library(magrittr)
library(readxl)
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(DataExplorer)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------
tidyverse 1.3.1 --
```

```
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x psych::%+%()       masks ggplot2::%+%()
## x psych::alpha()     masks ggplot2::alpha()
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

2)load file into a dataframe and data structure visualization

```
df_dodgers <- read.csv("C:/Users/goess/Downloads/dodgers.csv")
df_dodgers <- data.frame(df_dodgers)
head(df_dodgers,10)
```

```
##    month day attend day_of_week  opponent temp  skies day_night cap
shirt
## 1    APR  10  56000     Tuesday   Pirates   67 Clear         Day  NO
NO
## 2    APR  11  29729   Wednesday   Pirates   58 Cloudy      Night  NO
NO
## 3    APR  12  28328    Thursday   Pirates   57 Cloudy      Night  NO
NO
## 4    APR  13  31601      Friday    Padres   54 Cloudy      Night  NO
NO
## 5    APR  14  46549    Saturday    Padres   57 Cloudy      Night  NO
NO
## 6    APR  15  38359      Sunday    Padres   65 Clear         Day  NO
NO
## 7    APR  23  26376      Monday    Braves   60 Cloudy      Night  NO
NO
## 8    APR  24  44014     Tuesday    Braves   63 Cloudy      Night  NO
NO
## 9    APR  25  26345   Wednesday    Braves   64 Cloudy      Night  NO
NO
## 10   APR  27  44807      Friday Nationals   66 Clear       Night  NO
NO
##    fireworks bobblehead
## 1         NO         NO
## 2         NO         NO
## 3         NO         NO
## 4        YES         NO
## 5         NO         NO
## 6         NO         NO
## 7         NO         NO
## 8         NO         NO
## 9         NO         NO
## 10       YES         NO
```

```
# Data structure
str(df_dodgers)
```

```
## '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" ...
```

We have 81 observations of 12 variables.

3) summary statistics

```
Hmisc::describe(df_dodgers)

## df_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
## --------------------------------------------------------------------
------------
```

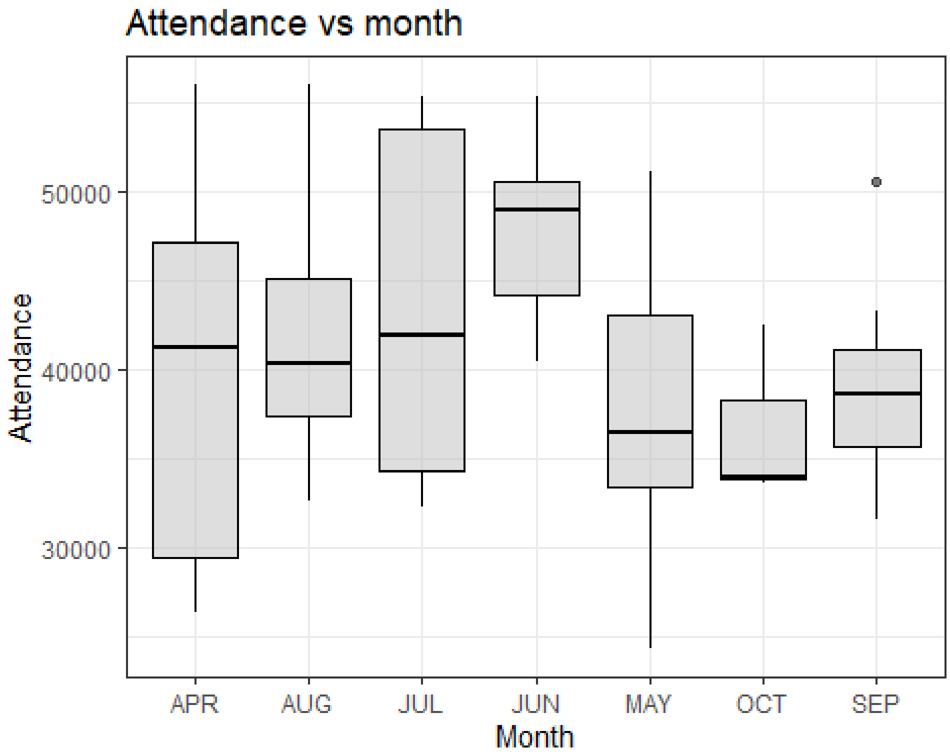
```
## day_of_week
##        n  missing distinct
##       81        0        7
##
## lowest : Friday    Monday    Saturday  Sunday    Thursday
## highest: Saturday  Sunday    Thursday  Tuesday   Wednesday
##
## Value         Friday    Monday  Saturday    Sunday  Thursday
Tuesday
## Frequency         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
## ---------------------------------------------------------------------
------------
```

The above summary tells us that the season takes place between April and November, that the matches can be played any day of the week (Wednesday to Sunday being the most frequent), day or night, with a preference for the night. The maximum attendance was 56,000 spectators. The promotions are Fireworks, cap, bobblehead and shirt.

I- Data visualization

4) let's visualize the data for better understanding let's plot attendance by day of the week, attendance by month , attendance by promotion, attendance by weather

```
#Attendance by month
ggplot2::ggplot(df_dodgers, ggplot2::aes(x=month, y=attend)) +
ggplot2::geom_boxplot(color="black", fill="grey", alpha=.5) +
ggplot2::labs(title="Attendance vs month", x='Month', y='Attendance') +
ggplot2:: theme_bw()
```

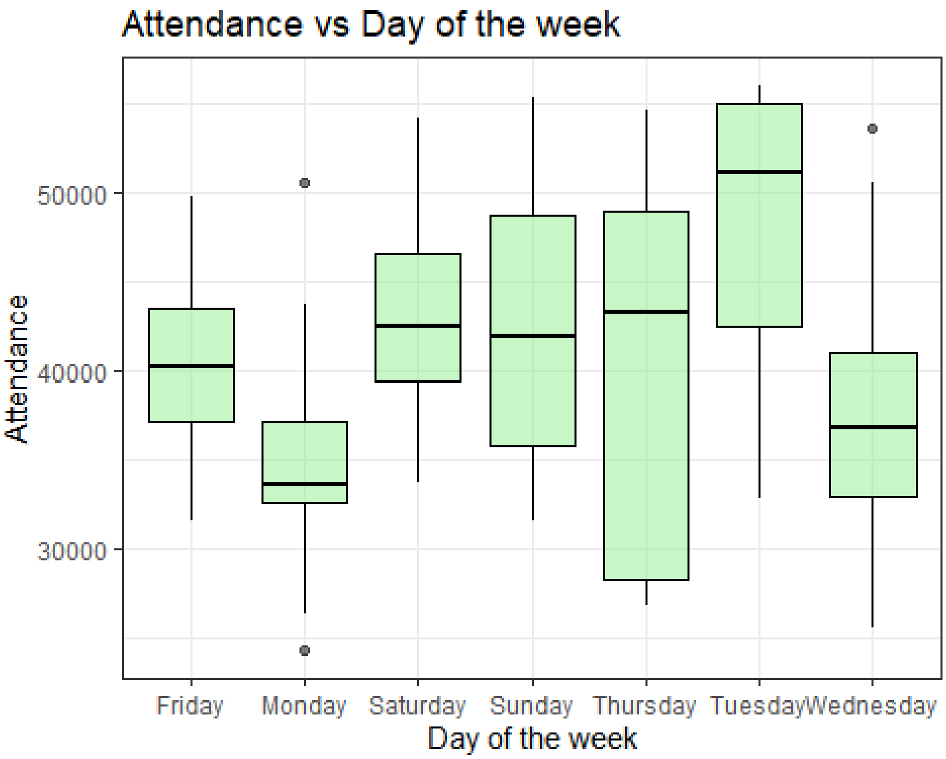


```
ggplot2::theme(plot.title = ggplot2::element_text(hjust = 1))

## List of 1
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

June is the most popular month of the season on average.

```
#Attendance vs day of the week
ggplot2::ggplot(df_dodgers, ggplot2::aes(x=day_of_week, y=attend)) +
ggplot2::geom_boxplot(color="black", fill="lightgreen", alpha=.5) +
ggplot2::labs(title="Attendance vs Day of the week", x='Day of the
week', y='Attendance') +
ggplot2:: theme_bw()
```



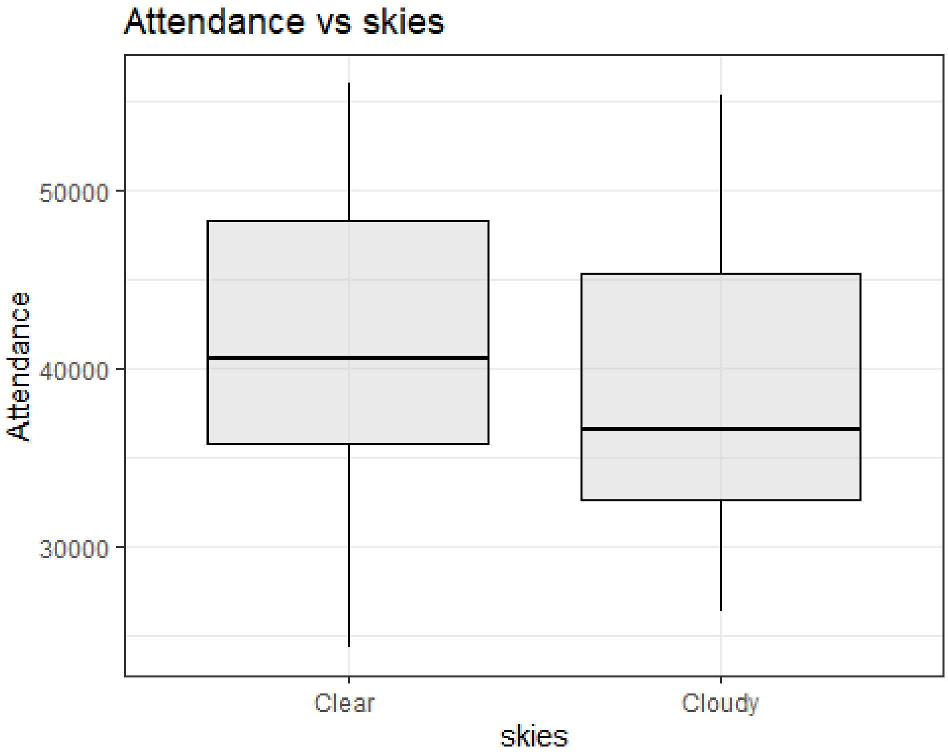
```
ggplot2::theme(plot.title = ggplot2::element_text(hjust = 1))

## List of 1
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

Tuesday is by far the day of the week with the most attendance in average followed by Thursday.

Weather and time of day can influence the expectation.

```
#attendance vs skies
ggplot2::ggplot(df_dodgers, ggplot2::aes(x=skies, y=attend)) +
ggplot2::geom_boxplot(color="black", fill="lightgrey", alpha=.5) +
ggplot2::labs(title="Attendance vs skies", x='skies', y='Attendance') +
ggplot2:: theme_bw()
```

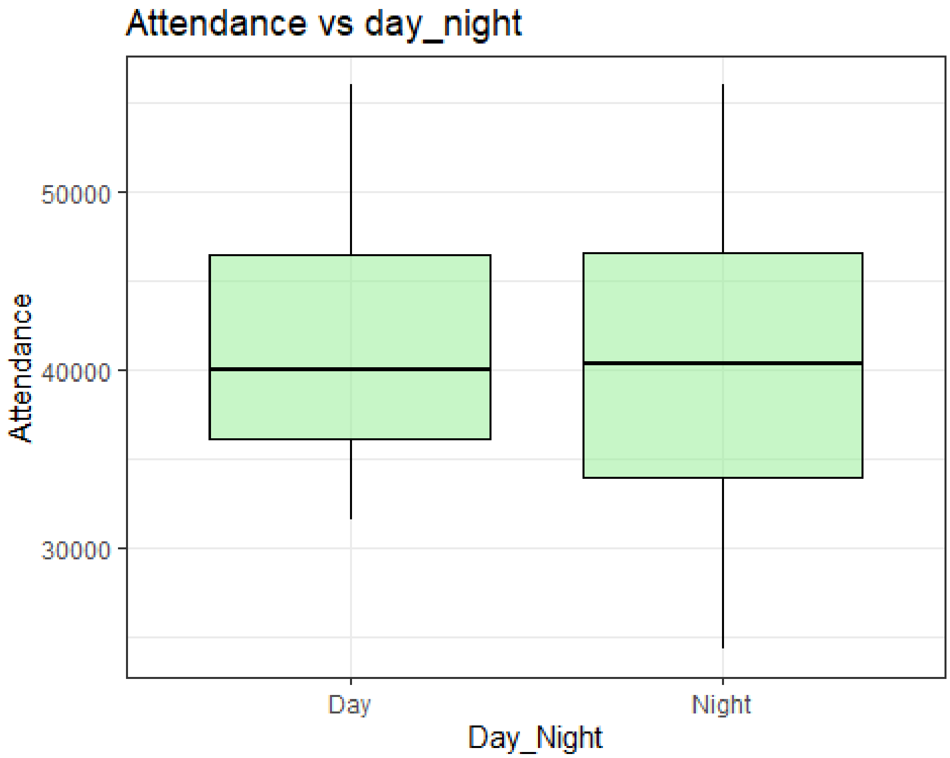


```
ggplot2::theme(plot.title = ggplot2::element_text(hjust = 1))
```

```
## List of 1
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

It is therefore obvious that clear skies draw the most crowds.

```
#attendance vs day_night
ggplot2::ggplot(df_dodgers, ggplot2::aes(x=day_night, y=attend)) +
ggplot2::geom_boxplot(color="black", fill="lightgreen", alpha=.5) +
ggplot2::labs(title="Attendance vs day_night", x='Day_Night',
y='Attendance') +
ggplot2:: theme_bw()
```



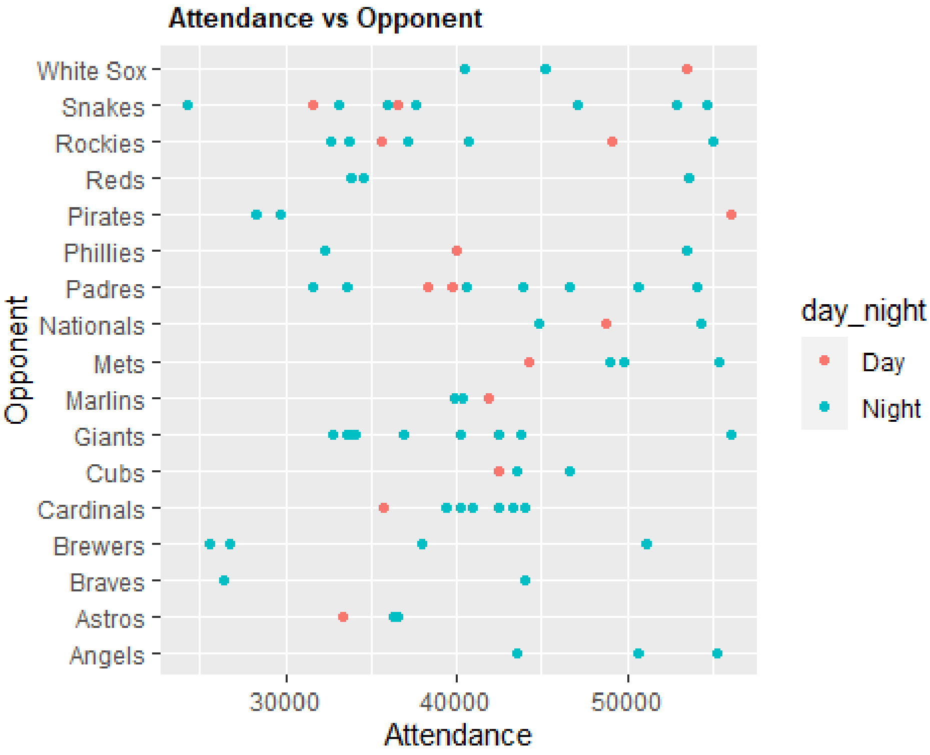
```
ggplot2::theme(plot.title = ggplot2::element_text(hjust = 1))
```

```
## List of 1
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

The attendance is very slightly higher at night games compared to day games.

5) Let's visualize attendance vs opponent

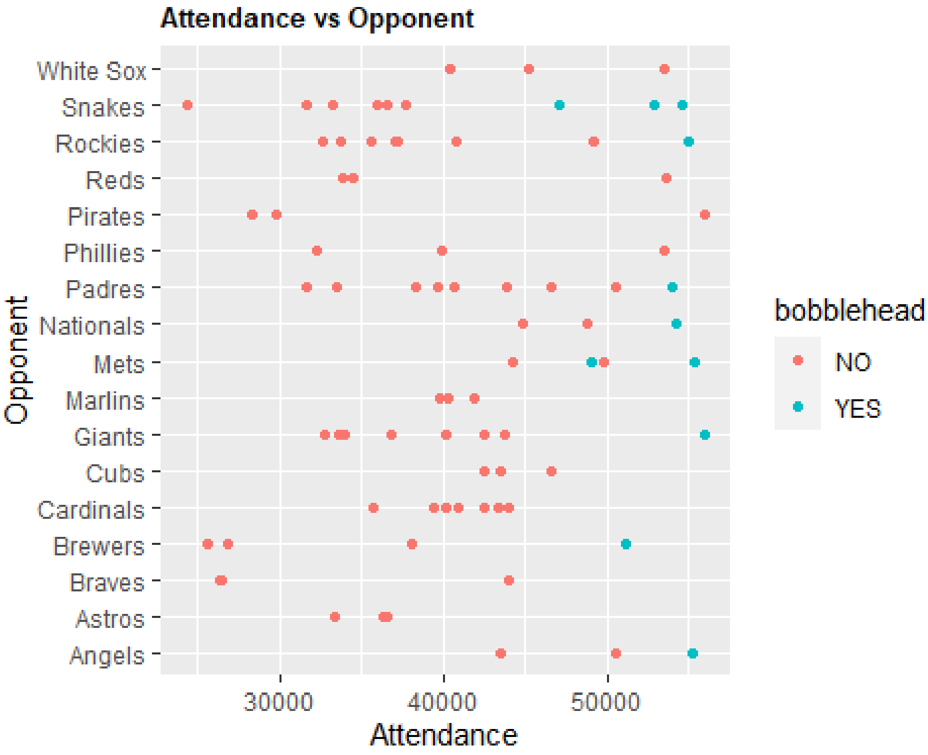
```
ggplot(df_dodgers, aes(x=attend, y=opponent, color=day_night)) +
        geom_point() +
        ggtitle(" Attendance vs Opponent") +
        theme(plot.title = element_text(lineheight=3, face="bold",
color="black", size=10)) +
        xlab("Attendance") +
        ylab("Opponent")
```



We can see that the matches against opponents from large metropolises attract bigger crowds most of the time in evening.

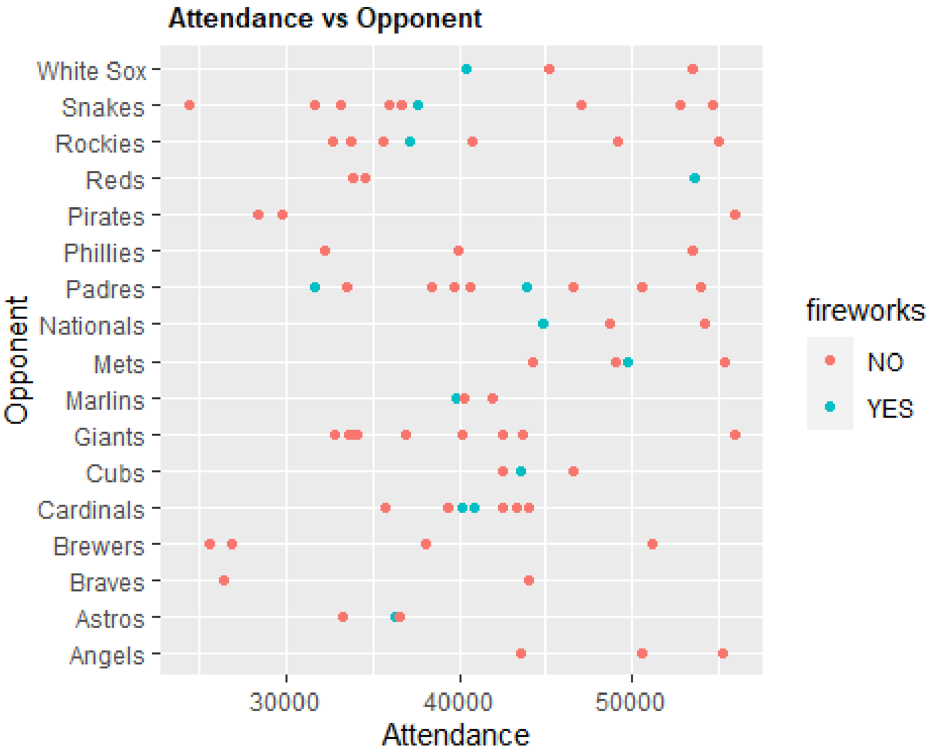
6) Let's combine attendance, weather, and time of the day and promotions.

```
#fireworks promotion
ggplot(df_dodgers, aes(x=attend, y=opponent, color=bobblehead)) +
        geom_point() +
        ggtitle("Attendance vs Opponent") +
        theme(plot.title = element_text(lineheight=3, face="bold",
color="black", size=10)) +
        xlab("Attendance") +
        ylab("Opponent")
```



The promotion bobblehead seems to draw more crowds to the stadium on match days compared to others.

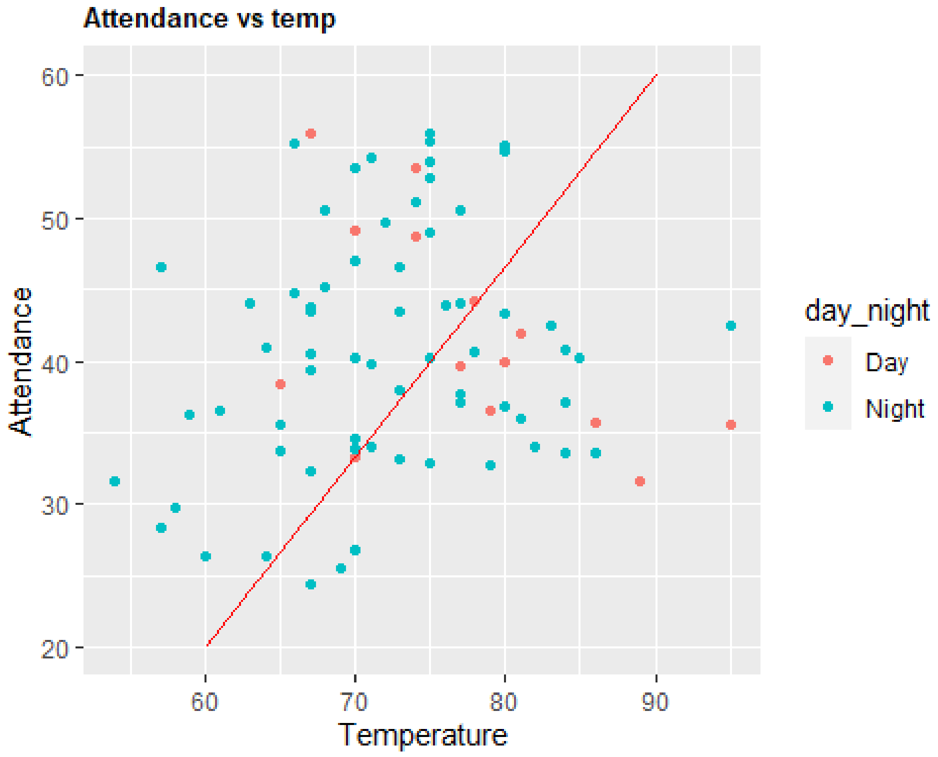
```
ggplot(df_dodgers, aes(x=attend, y=opponent, color=fireworks)) +
        geom_point() +
        ggtitle(" Attendance vs Opponent") +
        theme(plot.title = element_text(lineheight=3, face="bold",
color="black", size=10)) +
        xlab("Attendance ") +
        ylab("Opponent")
```



Fireworks promotion seems to have a limited impact on crowds.

Let's visualize atttendance vs temperature.

```
ggplot(df_dodgers, aes(x=temp, y=attend/1000, color=day_night)) +
        geom_point() +
        geom_line(data = data.frame(x = c(60,90), y = c(20,60)), aes(x
= x, y = y), colour = "red")+
        ggtitle("Attendance vs temp") +
        theme(plot.title = element_text(lineheight=3, face="bold",
color="black", size=10)) +
        ylab("Attendance") +
        xlab("Temperature")
```



The attendance in the stadium peaks when temperatures are between 65F and 80F.

The variables retained for the regression are, therefore: skies, bobblehead, temperature, opponents, day of the week and month.

II- Regression model

```
lmattendance = lm(attend~skies + bobblehead + temp + opponent +
day_of_week + month,data = df_dodgers)
summary(lmattendance)

##
## Call:
## lm(formula = attend ~ skies + bobblehead + temp + opponent +
##     day_of_week + month, data = df_dodgers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10186.8  -3174.7   -458.8   2563.0  13361.4
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           40319.57   14540.04   2.773  0.00783 **
## skiesCloudy           -1889.22    2395.75  -0.789  0.43416
## bobbleheadYES         10213.06    3141.63   3.251  0.00208 **
## temp                     39.55     249.78   0.158  0.87485
## opponentAstros        -8253.86   11337.07  -0.728  0.47005
## opponentBraves        -9088.01   12106.90  -0.751  0.45645
## opponentBrewers      -11362.42   11794.77  -0.963  0.34011
## opponentCardinals     -4228.43   11350.27  -0.373  0.71110
## opponentCubs          -4240.44   11722.39  -0.362  0.71910
## opponentGiants        -7829.80   11063.75  -0.708  0.48248
## opponentMarlins       -8319.23   11531.49  -0.721  0.47407
## opponentMets          -2499.07    6311.33  -0.396  0.69385
## opponentNationals      1899.71   12584.40   0.151  0.88063
## opponentPadres        -3406.07   10184.70  -0.334  0.73948
## opponentPhillies      -4560.05   10626.85  -0.429  0.66973
## opponentPirates       -4245.42   12215.79  -0.348  0.72968
## opponentReds          -5715.77   10450.90  -0.547  0.58692
## opponentRockies       -7825.28   10933.88  -0.716  0.47758
## opponentSnakes       -10580.21   10522.75  -1.005  0.31961
## opponentWhite Sox     -1150.84    5927.23  -0.194  0.84685
## day_of_weekMonday     -3173.52    3426.53  -0.926  0.35890
## day_of_weekSaturday    1696.30    2652.28   0.640  0.52544
## day_of_weekSunday       890.31    3431.38   0.259  0.79637
## day_of_weekThursday   -2070.79    4145.03  -0.500  0.61960
## day_of_weekTuesday     5118.79    3823.31   1.339  0.18680
## day_of_weekWednesday   -632.62    3617.53  -0.175  0.86190
## monthAUG               4876.29    7935.47   0.614  0.54173
## monthJUL               3469.08    6455.00   0.537  0.59341
## monthJUN               3595.93   11271.51   0.319  0.75106
## monthMAY               1210.24    6173.31   0.196  0.84539
## monthOCT               1097.30    9361.41   0.117  0.90717
## monthSEP               1173.16    7795.62   0.150  0.88100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6374 on 49 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.4099
## F-statistic: 2.793 on 31 and 49 DF,  p-value: 0.0006339
```

Results Interpretation, First of all, the value of the adjusted R-squared is 0.41. This means that the independent variables chosen together contribute to explaining only 40% of the variability of the attendance. Second, by observing the p-value of the different variables, The p-value for bobblehead yes is 0.00208. A small value means that age is probably a good addition to my model. Going through the p-values of each variable and choosing the smallest, we can consider 0.19 for the day of the week, Tuesday, 0.54 for August and 0.34 for opponent brewers. Third, let us look at each of the variables' estimate or correlation coefficient. A positive coefficient indicates that as the value of the independent variable increases.

It should be noted that the sign is essential in this case. So let us find the variables with a high and positive correlation coefficient. The Estimate Bobblehead has a value of 10213, national opponent 1900, day of the week Tuesday 5119 and 4876 for August.

Conclusion

A marketing promotion will have maximum impact if carried on a night where the bobblehead promotion is also carried out, preferably on a Tuesday and in June because June is the month with the highest attendance on average.