BA Homework4

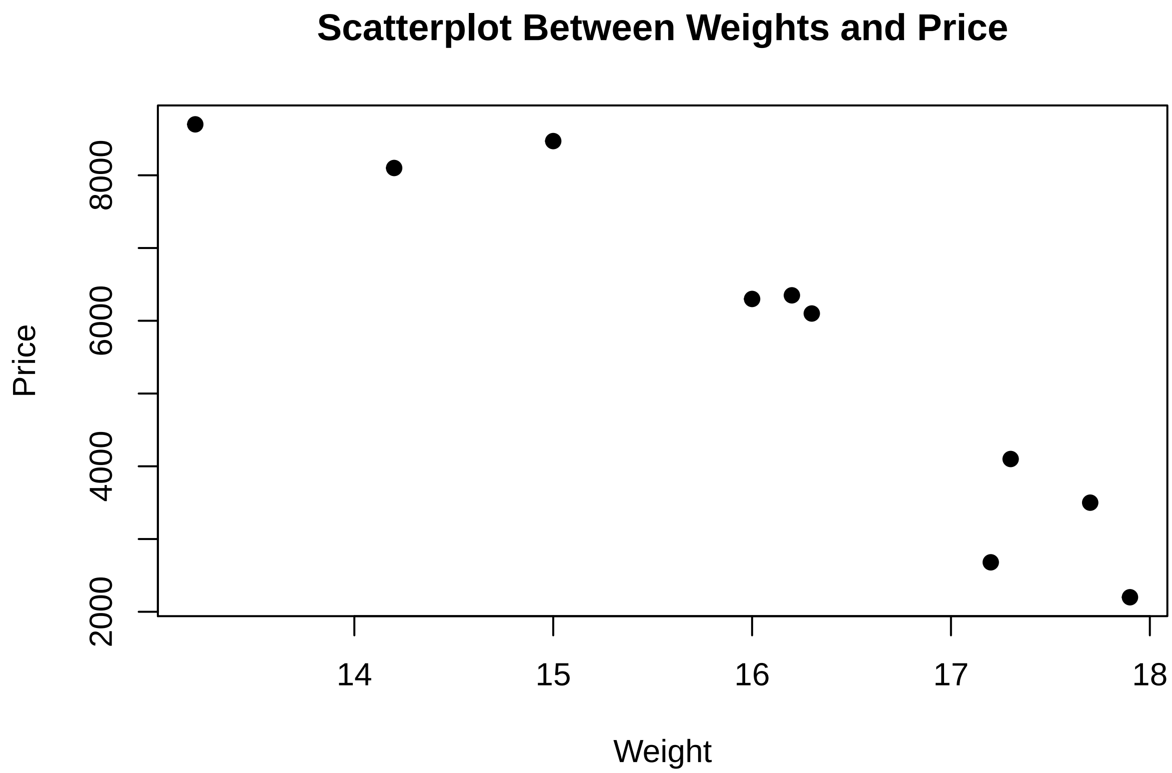
Prblem 1 create the data frame

```
df1 <- data.frame("Model" = c('Fieero 7B','HX 5000','Durbin Ultralight','Schmidt','WSilton Advanced','Bi
df1
```

```
##                Model Weight.lb. Price...
## 1          Fieero 7B       17.9     2200
## 2            HX 5000       16.2     6350
## 3  Durbin Ultralight       15.0     8470
## 4            Schmidt       16.0     6300
## 5   WSilton Advanced       17.3     4100
## 6    Bicyclette velo       13.2     8700
## 7       Supremo Team       16.3     6100
## 8          XTC Racer       17.2     2680
## 9       DOnofrio Pro       17.7     3500
## 10      Americana #6       14.2     8100
```

a. scatter chart between weights and price

```
plot(df1$Weight.lb., df1$Price..., main="Scatterplot Between Weights and Price",
   xlab="Weight", ylab="Price", pch=19)
```



Scatter plot shows there is a negative linear relationship between weights and price.

b.Estimated regression model

```
lm1 <- lm(df1$Price...~df1$Weight.lb.)
summary(lm1)
```

```
##
## Call:
## lm(formula = df1$Price... ~ df1$Weight.lb.)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1387.1  -715.9   164.6   679.9  1237.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     28818.0     3267.3   8.820 2.15e-05 ***
## df1$Weight.lb.  -1439.0      202.1  -7.121 9.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 942.3 on 8 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8467
## F-statistic:  50.7 on 1 and 8 DF,  p-value: 9.994e-05
```

let price be the y, weight be the x y=28818.0-1439.0*x c.From the summary table we can see that p-values for beta0 and beta1 is less than 0.05 and we have '***' for both parameters which means beta0 and beta1 are significant and not equal to zero at 0.05 level of significance.

d.From the Mutiple R-squared:0.8637, we can say the answer would be 86.37%

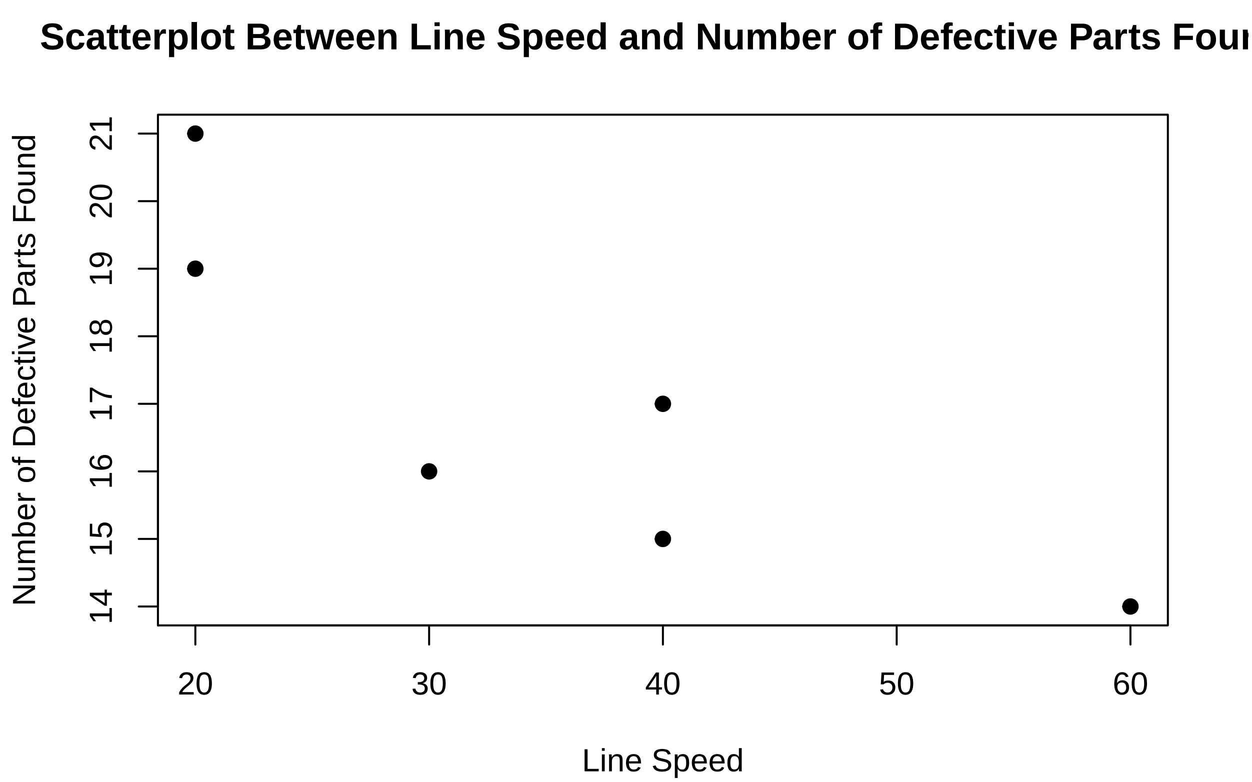
Problem2 create the data frame

```
df2 <- data.frame("Line Speed(ft/min)" = c(20,20,40,30,60,40), "Number of Defective Parts Found" = c(21
df2
```

```
##   Line.Speed.ft.min. Number.of.Defective.Parts.Found
## 1                 20                              21
## 2                 20                              19
## 3                 40                              15
## 4                 30                              16
## 5                 60                              14
## 6                 40                              17
```

a.scatter chart

```
plot(df2$Line.Speed.ft.min.,df2$Number.of.Defective.Parts.Found,
     main="Scatterplot Between Line Speed and Number of Defective Parts Found",
   xlab="Line Speed", ylab="Number of Defective Parts Found", pch=19)
```



Scatter plot shows there is a negative linear relationship between line speed and number of defective parts found.

b.Estimated regression model

```r
lm2 <- lm(df2$Number.of.Defective.Parts.Found~df2$Line.Speed.ft.min.)
summary(lm2)
```

```
##
## Call:
## lm(formula = df2$Number.of.Defective.Parts.Found ~ df2$Line.Speed.ft.min.)
##
## Residuals:
##       1       2       3       4       5       6
##  1.7826 -0.2174 -1.2609 -1.7391  0.6957  0.7391
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            22.17391    1.65275  13.416 0.000179 ***
## df2$Line.Speed.ft.min. -0.14783    0.04391  -3.367 0.028135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.489 on 4 degrees of freedom
## Multiple R-squared:  0.7391, Adjusted R-squared:  0.6739
## F-statistic: 11.33 on 1 and 4 DF,  p-value: 0.02813
```

let number of defective parts found be y and let line speed be x y=22.17391-0.14783*x

c.From the summary table we can see that p-value for beta0 is less than 0.01 but p-value for beta1 is larger than 0.01.Thus we can say beta0 is significant and not equal to zero at 0.01 level of significance but beta1 is not significant and equal to zero at 0.01 level of significance.

d.From the Mutiple R-squared:0.7391, we can say the answer would be 73.91%

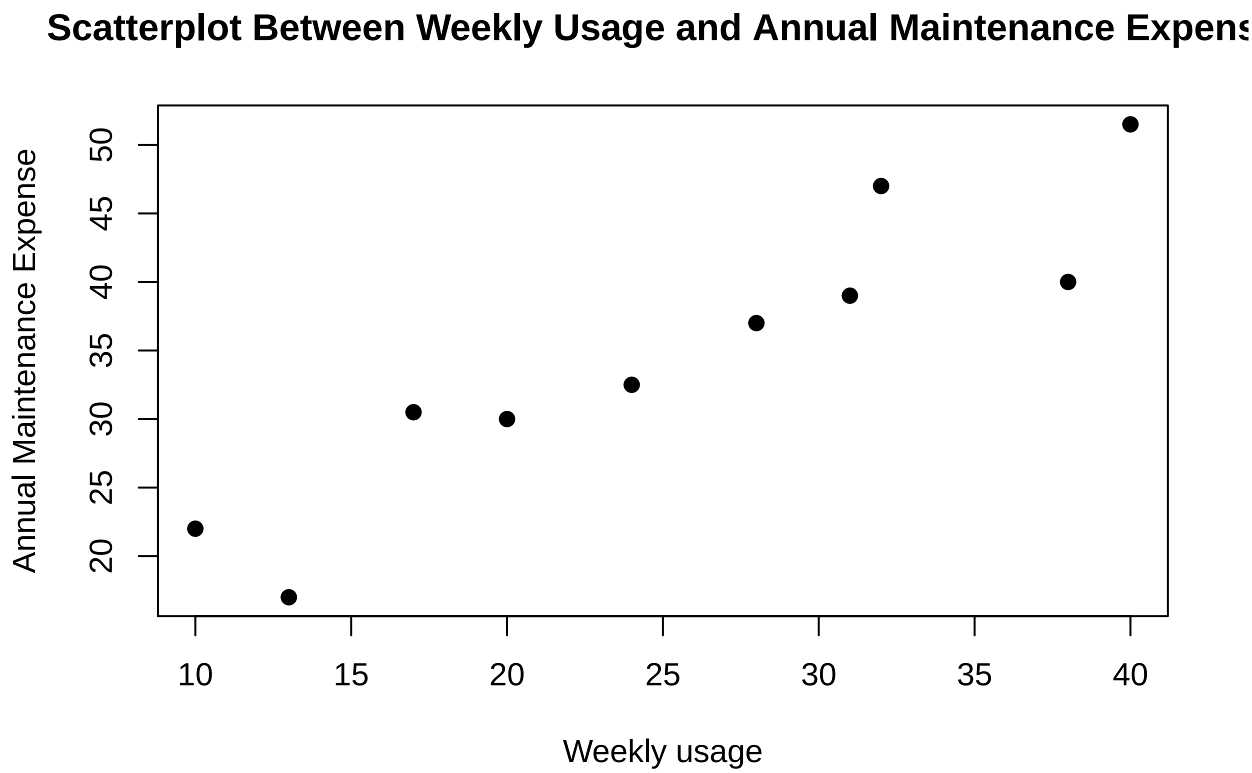
Problem3 create the data frame

```r
df3 <- data.frame("Weekly Usage(hours)" = c(13,10,20,28,32,17,24,31,40,38), "Annual Maintenance Expense
df3
```

```
##    Weekly.Usage.hours. Annual.Maintenance.Expense.hundreds.of.dollars.
## 1                   13                                            17.0
## 2                   10                                            22.0
## 3                   20                                            30.0
## 4                   28                                            37.0
## 5                   32                                            47.0
## 6                   17                                            30.5
## 7                   24                                            32.5
## 8                   31                                            39.0
## 9                   40                                            51.5
## 10                  38                                            40.0
```

a.scatter chart

```r
plot(df3$Weekly.Usage.hours.,df3$Annual.Maintenance.Expense.hundreds.of.dollars.,
     main="Scatterplot Between Weekly Usage and Annual Maintenance Expense",
   xlab="Weekly usage", ylab="Annual Maintenance Expense", pch=19)
```



Scatter plot shows there is a positive linear relationship between weekly usage and annual maintenance expense.

b.Estimated regression model

```
lm3 <- lm(df3$Annual.Maintenance.Expense.hundreds.of.dollars.~df3$Weekly.Usage.hours.)
summary(lm3)
```

```
##
## Call:
## lm(formula = df3$Annual.Maintenance.Expense.hundreds.of.dollars. ~
##     df3$Weekly.Usage.hours.)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7587 -1.0411  0.0895  2.6102  5.9619
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               10.5280     3.7449   2.811 0.022797 *
## df3$Weekly.Usage.hours.    0.9534     0.1382   6.901 0.000124 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.25 on 8 degrees of freedom
## Multiple R-squared:  0.8562, Adjusted R-squared:  0.8382
## F-statistic: 47.62 on 1 and 8 DF,  p-value: 0.0001244
```

let weekly usage be x and annual maintenance expense be y. y=10.5280+0.9534*x

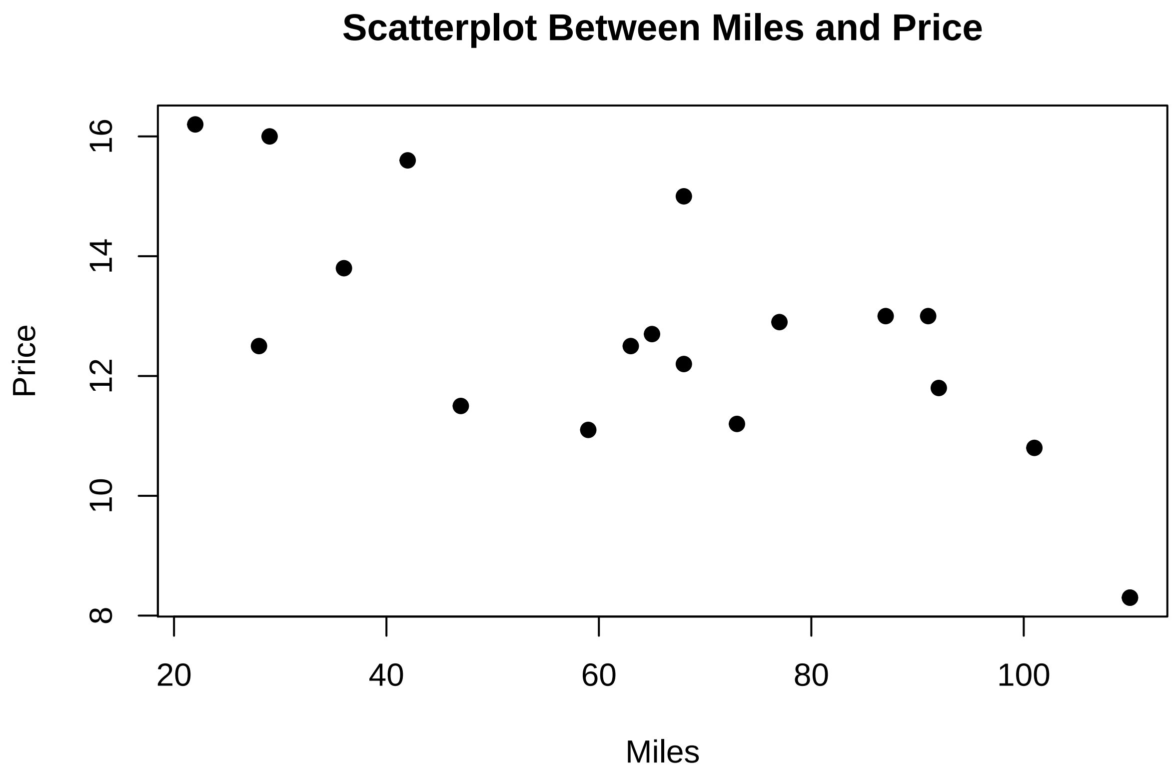
Problem4 create the data frame

```
df4 <- data.frame("Miles(1000s)" = c(22,29,36,47,63,77,73,87,92,101,110,28,59,68,68,91,42,65,110), "Pri
df4
```

```
##    Miles.1000s. Price..1000s.
## 1            22          16.2
## 2            29          16.0
## 3            36          13.8
## 4            47          11.5
## 5            63          12.5
## 6            77          12.9
## 7            73          11.2
## 8            87          13.0
## 9            92          11.8
## 10          101          10.8
## 11          110           8.3
## 12           28          12.5
## 13           59          11.1
## 14           68          15.0
## 15           68          12.2
## 16           91          13.0
## 17           42          15.6
## 18           65          12.7
## 19          110           8.3
```

a.scatter chart

```
plot(df4$Miles.1000s., df4$Price..1000s.,
     main="Scatterplot Between Miles and Price",
   xlab="Miles", ylab="Price", pch=19)
```



Scatter plot shows there is a negative linear relationship between miles and price.

b.Estimated regression model

```
lm4 <- lm(df4$Price..1000s.~df4$Miles.1000s.)
summary(lm4)
```

```
##
## Call:
## lm(formula = df4$Price..1000s. ~ df4$Miles.1000s.)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32408 -1.34194  0.05055  1.12898  2.52687
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      16.46976    0.94876  17.359 2.99e-12 ***
## df4$Miles.1000s. -0.05877    0.01319  -4.455 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.541 on 17 degrees of freedom
## Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
## F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

let miles be x and price be y. y=16.46976-0.05877*x c.From the summary table we can see that p-values for beta0 and beta1 is less than 0.01 and we have '***' for both parameters which means beta0 and beta1 are significant and not equal to zero at 0.05 level of significance.

d.From the Mutiple R-squared:0.5387, we can say the answer would be 53.87%

e.

```
pred4<-predict(lm4)
residual4<-residuals(lm4)
df4_e<-data.frame("Observation" = 1:19, "Price Prediction" = pred4 , 'Residuals' = residual4, stringsAsI
df4_e
```

```
##    Observation Price.Prediction   Residuals
## 1            1         15.17673  1.02327147
## 2            2         14.76531  1.23468899
## 3            3         14.35389 -0.55389349
## 4            4         13.70738 -2.20738023
## 5            5         12.76700 -0.26699732
## 6            6         11.94416  0.95583772
## 7            7         12.17926 -0.97925801
## 8            8         11.35642  1.64357704
## 9            9         11.06255  0.73744670
## 10          10         10.53359  0.26641209
## 11          11         10.00462 -1.70462253
## 12          12         14.82408 -2.32408494
## 13          13         13.00209 -1.90209305
## 14          14         12.47313  2.52687234
## 15          15         12.47313 -0.27312766
## 16          16         11.12133  1.87867277
## 17          17         14.00125  1.59875011
## 18          18         12.64945  0.05055054
## 19          19         10.00462 -1.70462253
```

```
#find the two smallest residuals in the Observations
sort(df4_e$Residuals)[1]
```

```
## [1] -2.324085
```

```
sort(df4_e$Residuals)[2]
```

```
## [1] -2.20738
```

The biggest bargains means they have highest negative residuals. Observation 4 which has -2.207 is the second bargain and Observation 12 which has -2.324 is the first bargain.

f.

```
x=60
y=16.46976-0.05877*x
y
```

```
## [1] 12.94356
```

Based on the estimated regression model, given 60000 miles, the predicted price will be $12943.56 We still need to combine other real factors to give a better final price but this price is good enough to offer if we only consider the miles.

Problem5 read and check data

```r
url<-'https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/dodgers.csv'
df5<-read.csv(url,header=TRUE,stringsAsFactors = TRUE)
head(df5)
```

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

```r
dim(df5)
```

```
## [1] 81 12
```

```r
names(df5)
```

```
##  [1] "month"       "day"         "attend"      "day_of_week" "opponent"
##  [6] "temp"        "skies"       "day_night"   "cap"         "shirt"
## [11] "fireworks"   "bobblehead"
```

1. Complete an exploratory data analysis and answer the following:

a. How many times did promotions take place during the year (cap vs shirts vs bobblehead vs fireworks)?

```r
table(df5$cap)
```

```
##
##  NO YES
##  79   2
```

2 times for cap

```r
table(df5$shirt)
```

```
##
##  NO YES
##  78   3
```

3 times for shirt

```r
table(df5$fireworks)
```

```
##
##  NO YES
##  67  14
```

14 times for fireworks

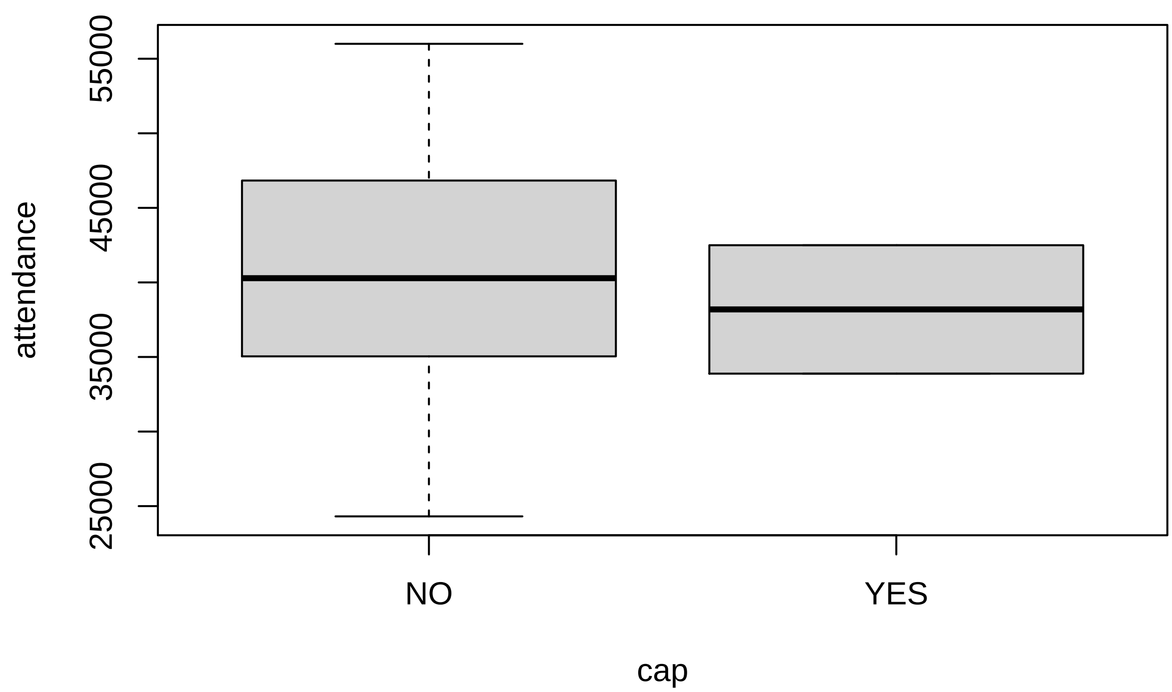
```r
table(df5$bobblehead)
```

```
##
##  NO YES
##  70  11
```

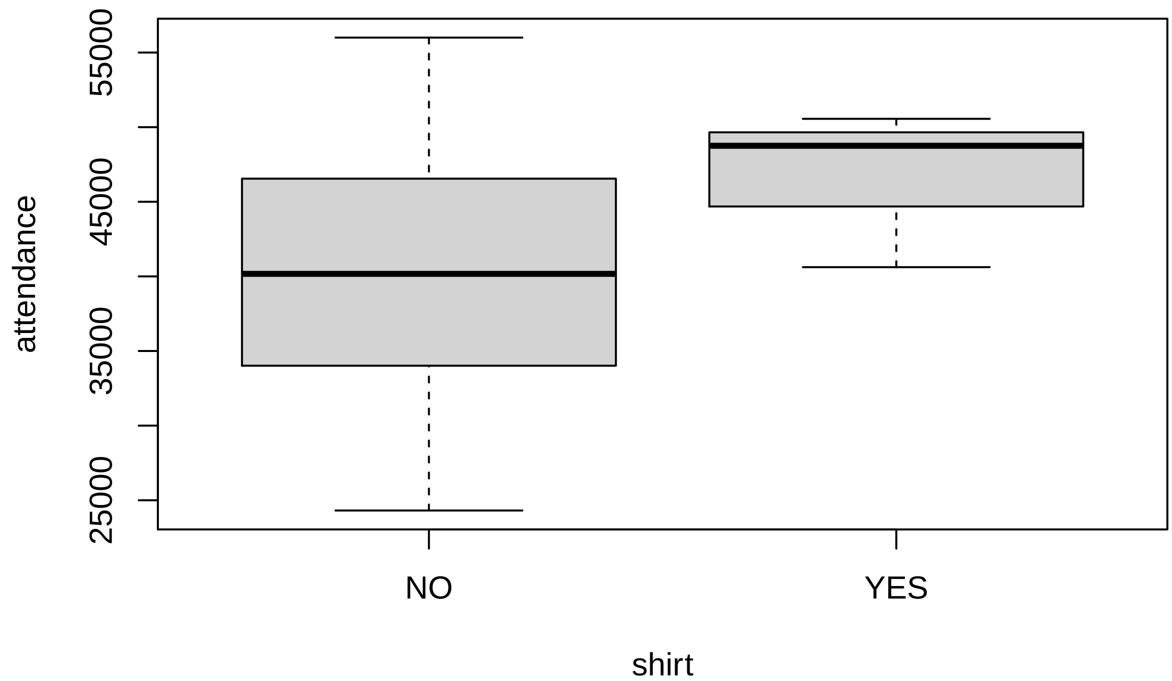
11 times for bobblehead

b.How does attendance vary with and without promotions

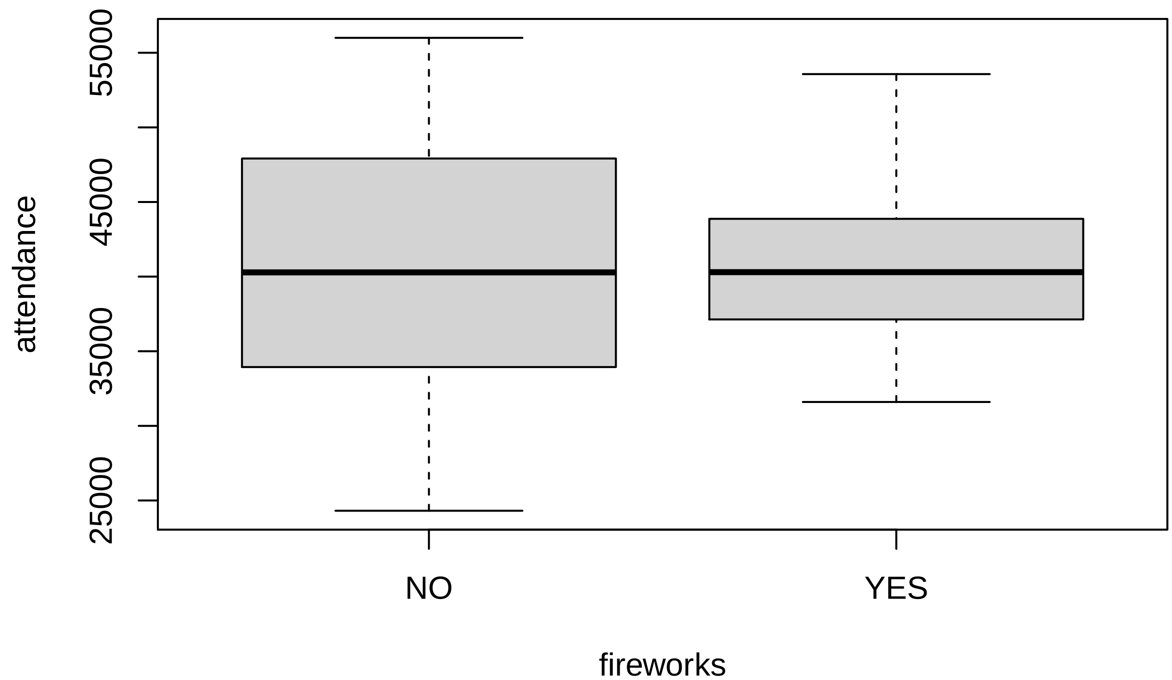
```r
plot(df5$cap,df5$attend,xlab='cap',ylab='attendance')
```



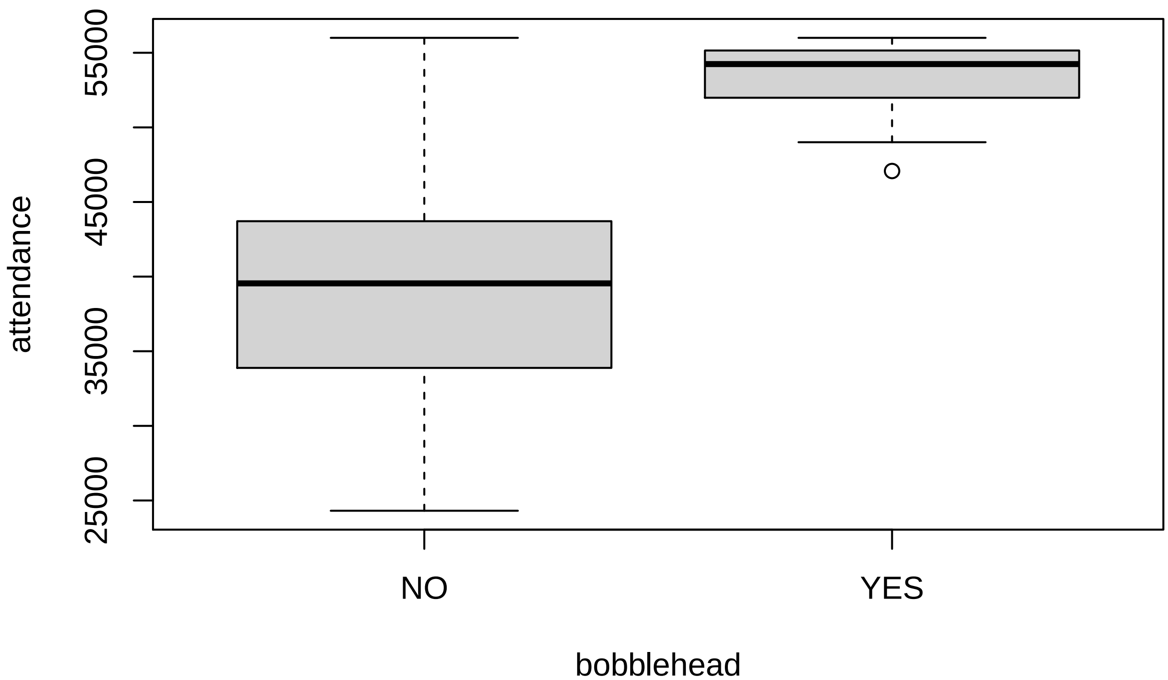
```r
plot(df5$shirt,df5$attend,xlab='shirt',ylab='attendance')
```




```r
plot(df5$fireworks,df5$attend,xlab='fireworks',ylab='attendance')
```



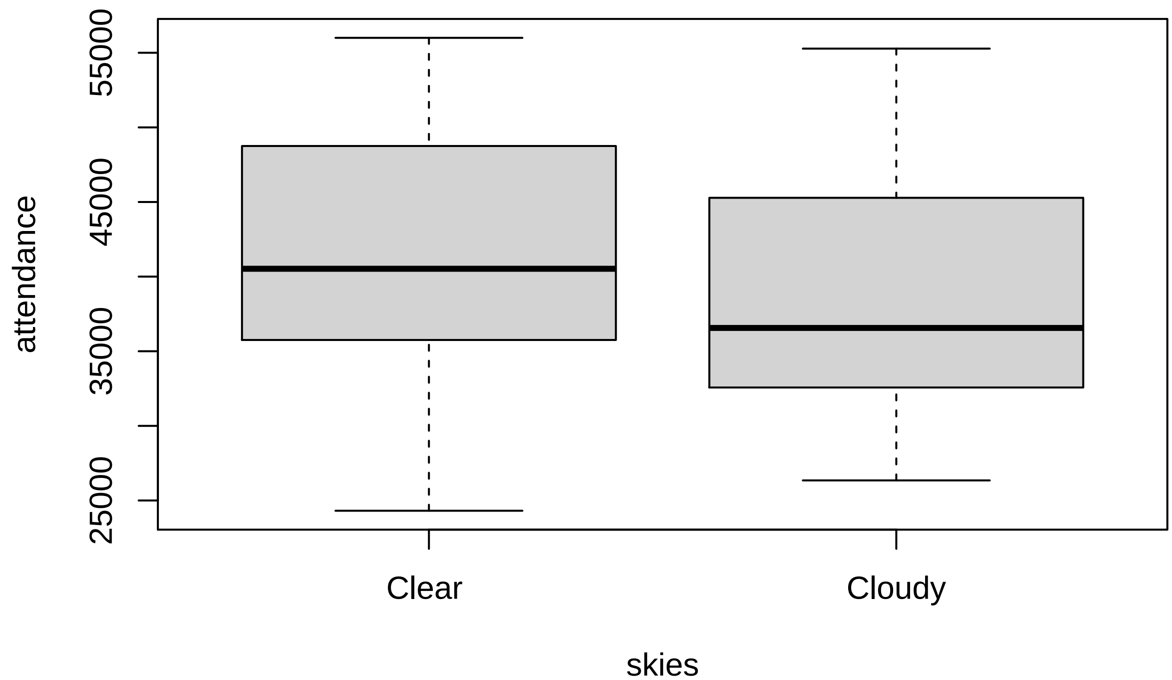

```
plot(df5$bobblehead,df5$attend,xlab='bobblehead',ylab='attendance')
```



In shirts and bobblehead, with promotions attendance will be higher. In cap and fireworks, promotions do no affect attendance too much.

c. What patterns exist with programming of games (weather, time, month, day, etc)? Attendance by weather

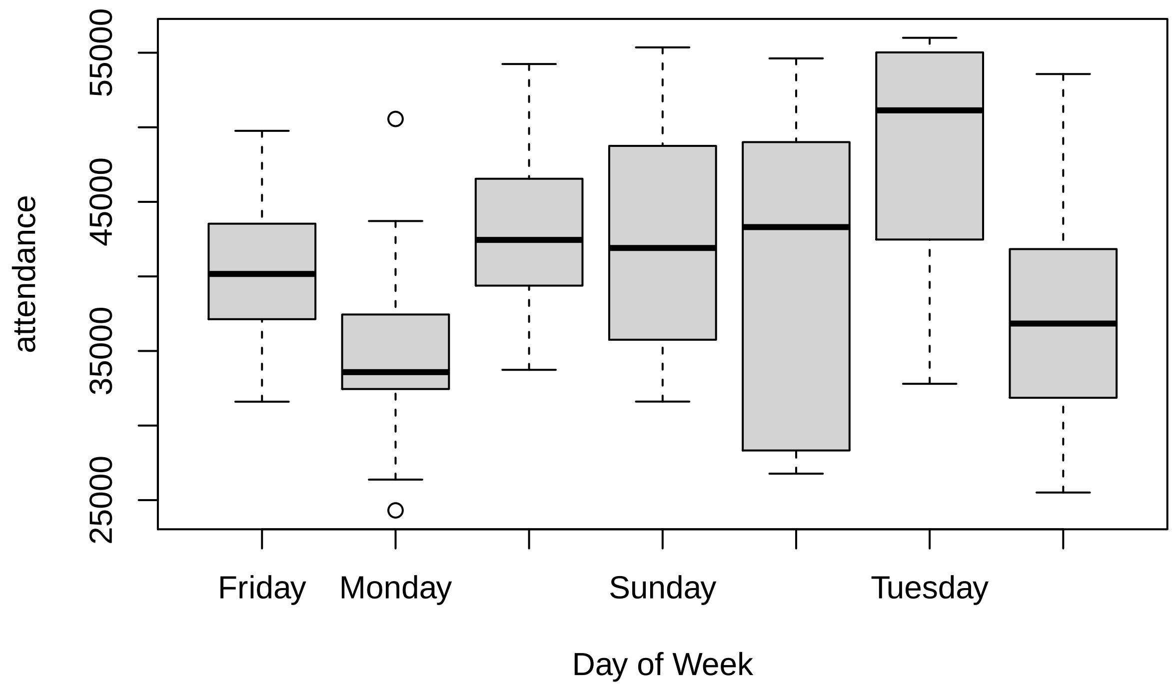
```
plot(df5$skies,df5$attend,xlab='skies',ylab='attendance')
```



Clear sky had higher attendance

Attendance by days

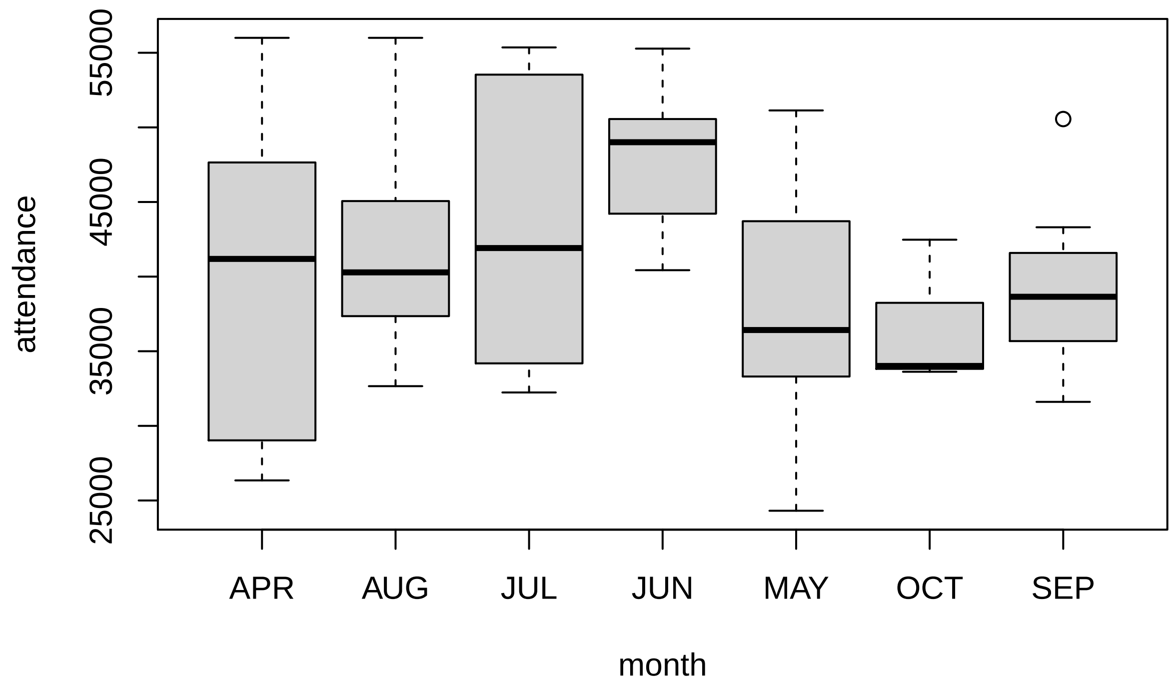
```r
plot(df5$day_of_week,df5$attend,xlab='Day of Week',ylab='attendance')
```



Tuesday had the highest overall attendance day in the week from the graph.

Attendance by month

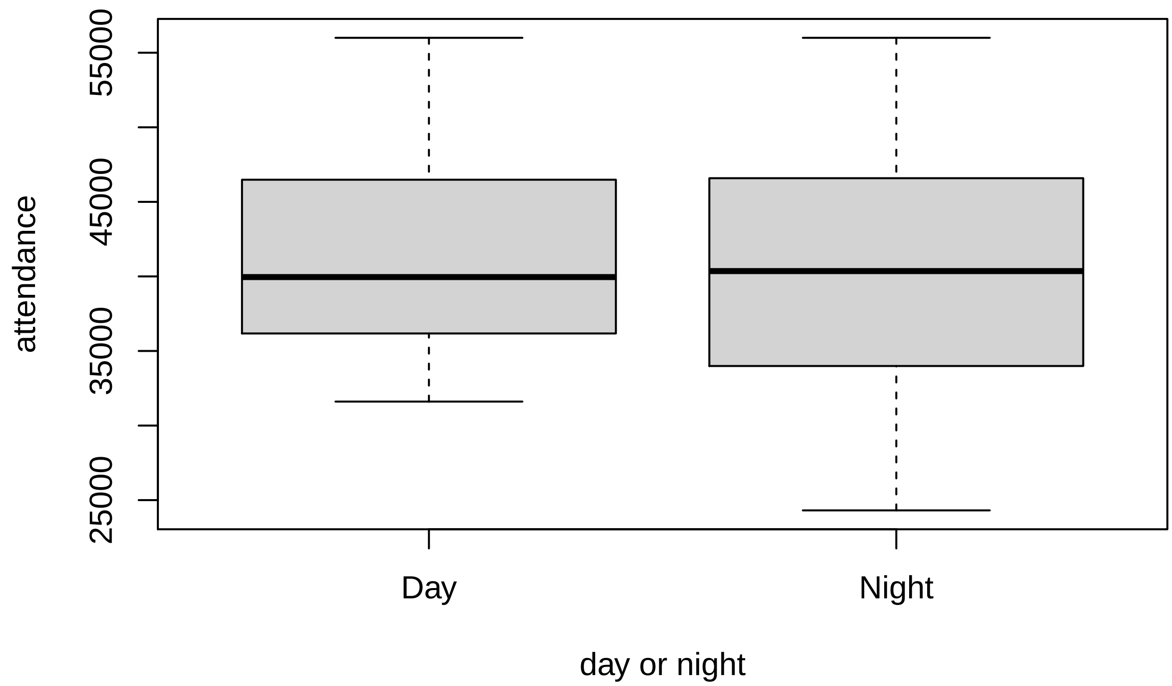
```
plot(df5$month,df5$attend,xlab='month',ylab='attendance')
```



June had the highest overall attendance day in the week from the graph.

Attendance by day_nights

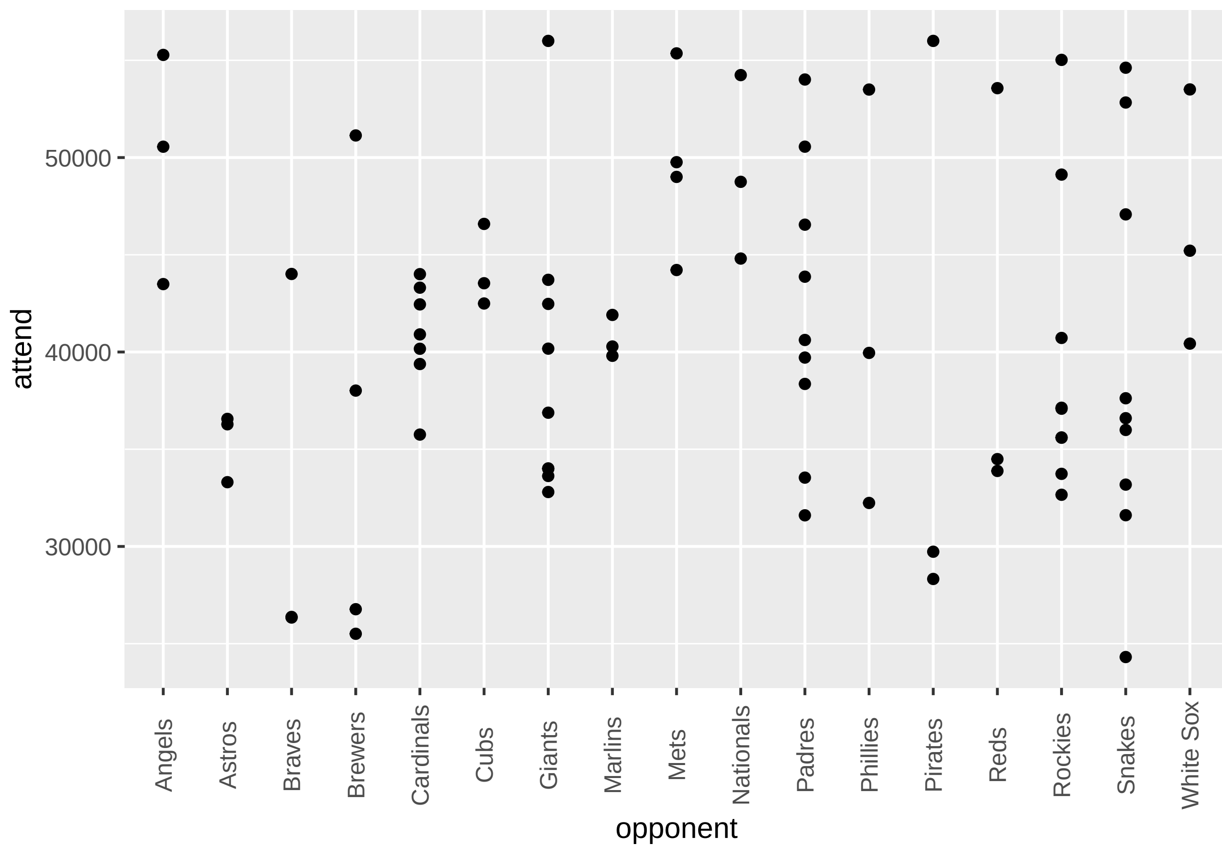
```
plot(df5$day_night,df5$attend,xlab='day or night',ylab='attendance')
```



There's not too much attendance difference between day and night.

d.Which opposing teams bring is attendance above average?

```
library(ggplot2)
ggplot(df5, aes(x = opponent, y = attend)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))
```



Opponents from large cities show higher attendance.

2. Answer the following questions using predictive modeling techniques:

a. Will the bobblehead promotions increase attendance?

```
lm5_ba<-lm(df5$attend~df5$bobblehead)
summary(lm5_ba)
```

```
##
## Call:
## lm(formula = df5$attend ~ df5$bobblehead)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14825.9  -5123.9    667.1   4171.1  16862.1
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         39137.9      811.6   48.22  < 2e-16 ***
## df5$bobbleheadYES  14006.7     2202.5    6.36 1.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6791 on 79 degrees of freedom
## Multiple R-squared:  0.3386, Adjusted R-squared:  0.3302
## F-statistic: 40.44 on 1 and 79 DF,  p-value: 1.217e-08
```

From problem 1 bobblehead vs attendance boxplot we can see that with bobblehead promotions, the attendance will be higher. We also can get the information from summary linear relationship between bobbleheadYES and attendance. We can see the pvalue for both beta0 and beta1 are very small and they have '***' which means both of them are significant and promotions for boblehead do affect the attendance.

b. Are bobblehead promotions better than all other promotions? put all the promotions into the model

```
lm5_all<-lm(df5$attend~df5$cap+df5$shirt+df5$fireworks+df5$bobblehead)
summary(lm5_all)
```

```
##
## Call:
## lm(formula = df5$attend ~ df5$cap + df5$shirt + df5$fireworks +
##     df5$bobblehead)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13889.1  -4466.1   -185.1   3729.1  17798.9
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       38201.08     933.09  40.940  < 2e-16 ***
## df5$capYES          -11.58    4803.39  -0.002   0.9981
## df5$shirtYES       8442.59    3958.77   2.133   0.0362 *
## df5$fireworksYES   2876.78    2010.56   1.431   0.1566
## df5$bobbleheadYES 14943.56    2215.26   6.746 2.64e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6664 on 76 degrees of freedom
## Multiple R-squared:  0.3873, Adjusted R-squared:  0.3551
## F-statistic: 12.01 on 4 and 76 DF,  p-value: 1.293e-07
```

From problem 1 4 different promotions vs attendance boxplots, we can say that the shirt and boblehead do affect the attendance. From the summary linear relationship, we can see that cap and firework have very high p-value and zero * which means they are not significant. shirt has higher p-value than bobblehead and only one * which means bobblehead is much more siginicant than shirt. Thus bobblehead promotions are better than all other promotions.

c. Giving your predictions, how many bobblehead should we ordered for the summer time (Jun - Aug)

```
summer<-subset(df5,df5$month =='JUN'|df5$month =='JUL'|df5$month =='AUG')#get the summer data
sum(summer$attend) #calculate the total attendance
```

```
## [1] 1599348
```

From question a, my estimated model between attendance and bobblehead is attendance = 39137.9+14006.7*bobblehead we have attendance now to calculate the bobblehead

```
bobblehead_pred = (sum(summer$attend)-39137.9)/14006.7
bobblehead_pred
```

```
## [1] 111.3903
```

About 112 bobblehead promotions needed for the summer time.