midterm_project

Ningze Zu

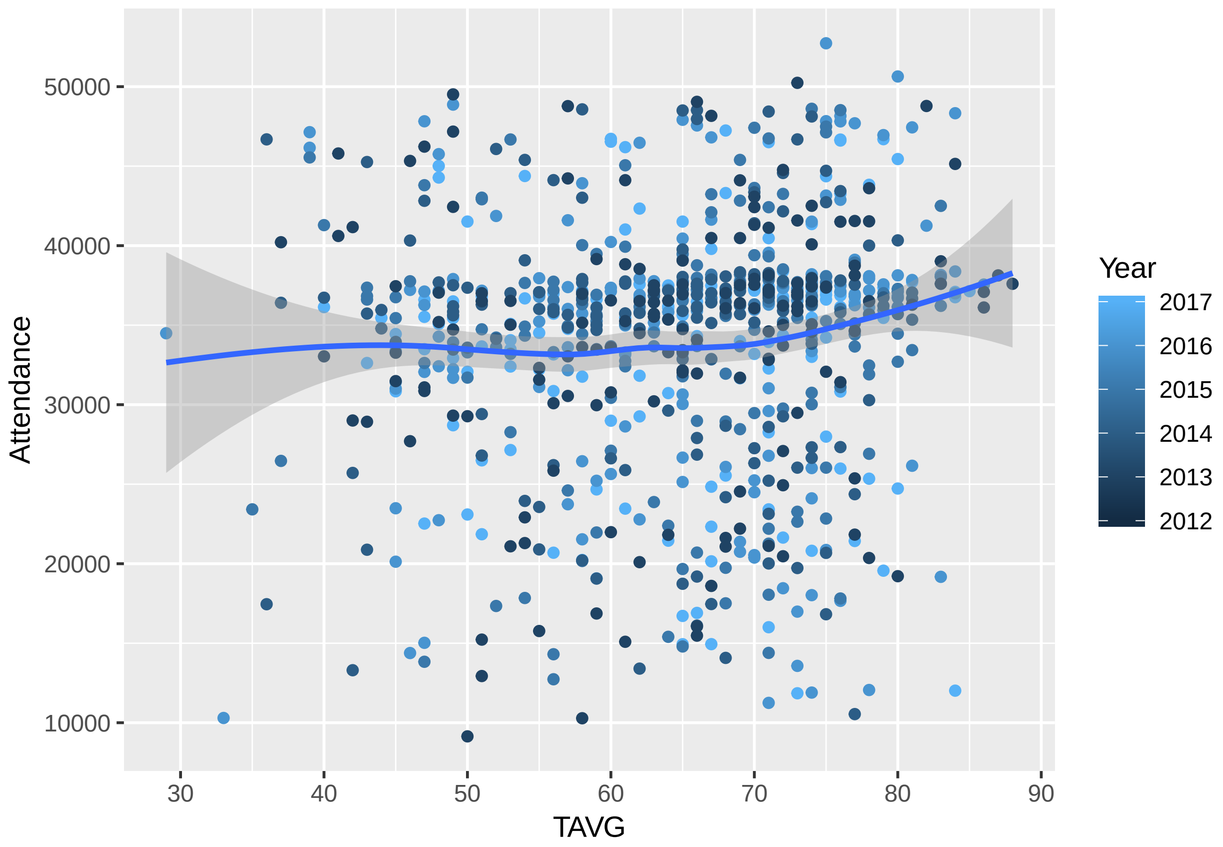
```
baseball <- read.csv("baseball_weather.csv", header = T)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

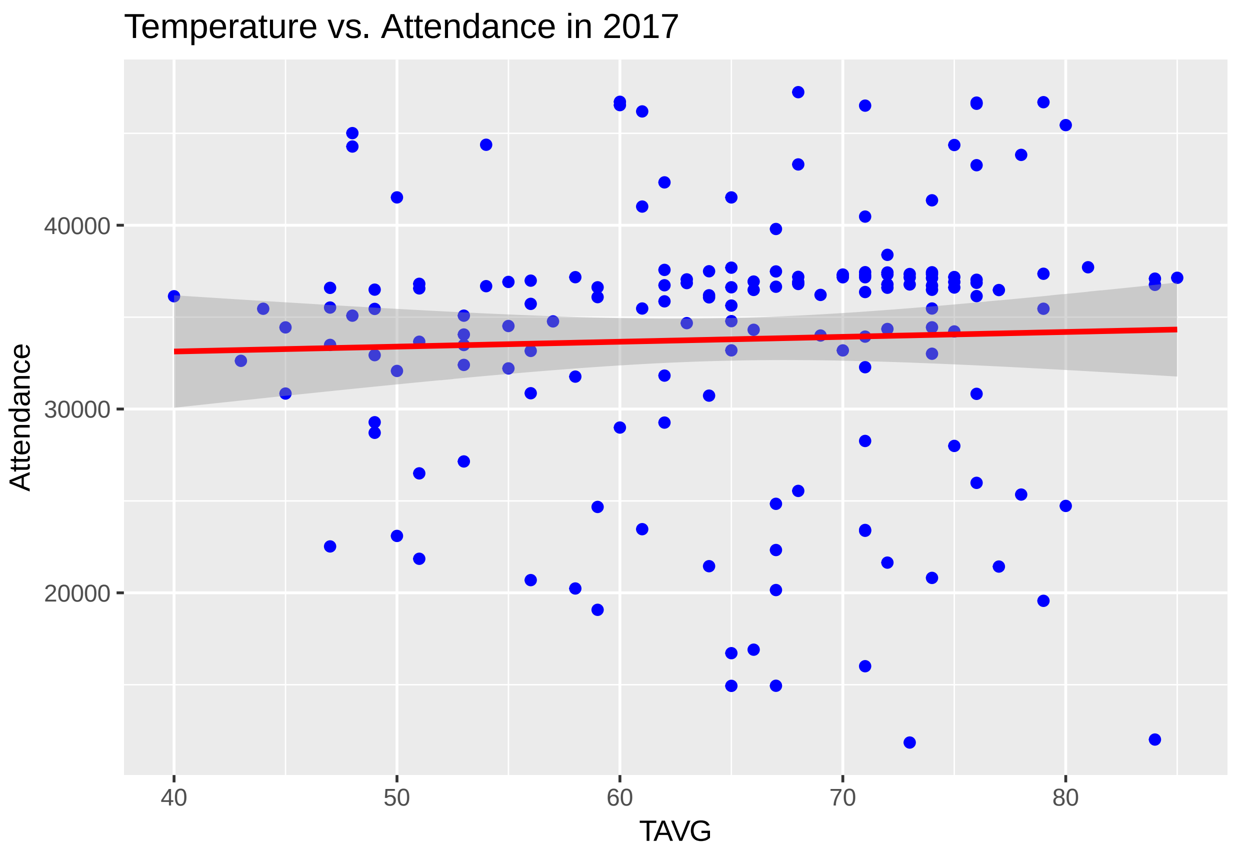
```
library(ggplot2)
library(knitr)
```

```
# Relationship between average temperature with attendance of 6 seasons
ggplot(baseball, mapping = aes(x = TAVG, y = Attendance)) +
  geom_point(mapping = aes(color = Year)) +
  geom_smooth()
```

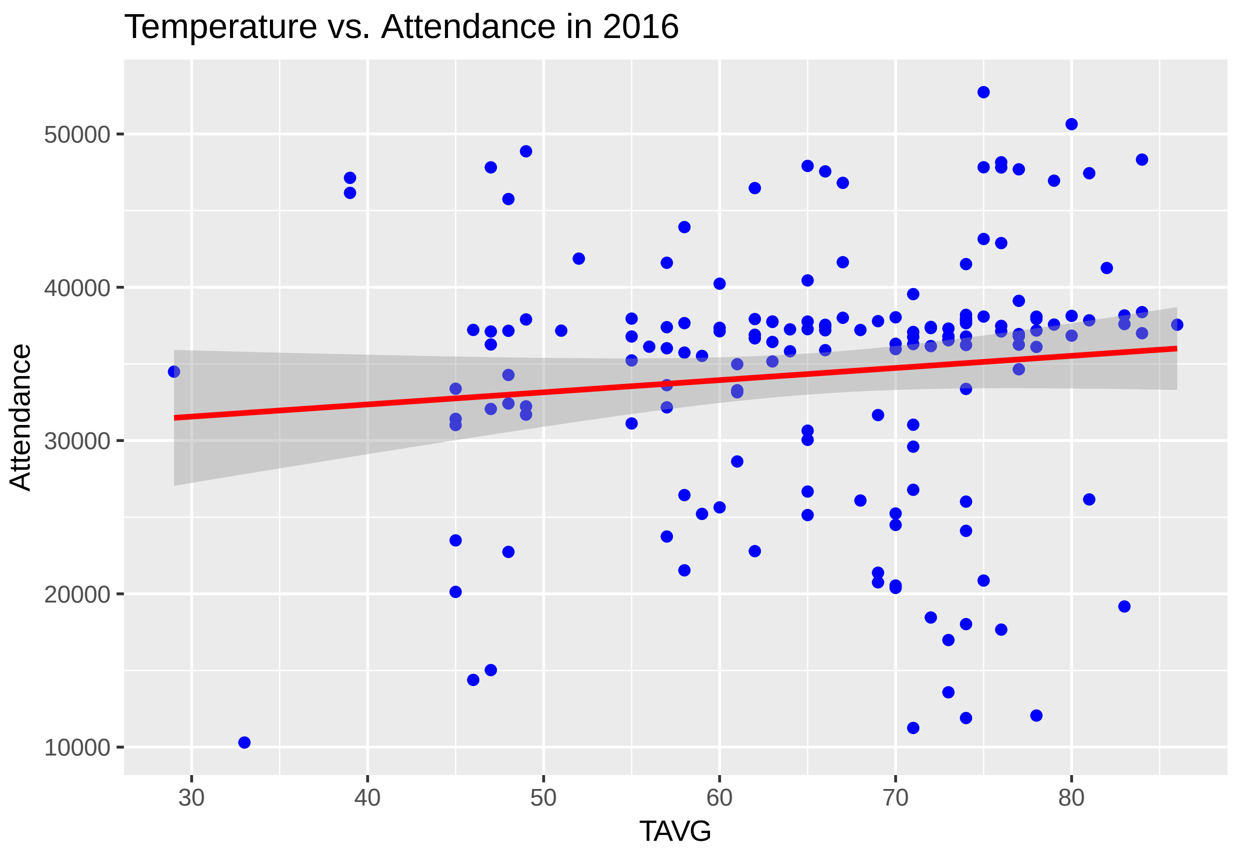
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 162 rows containing non-finite values (stat_smooth).

## Warning: Removed 162 rows containing missing values (geom_point).
```



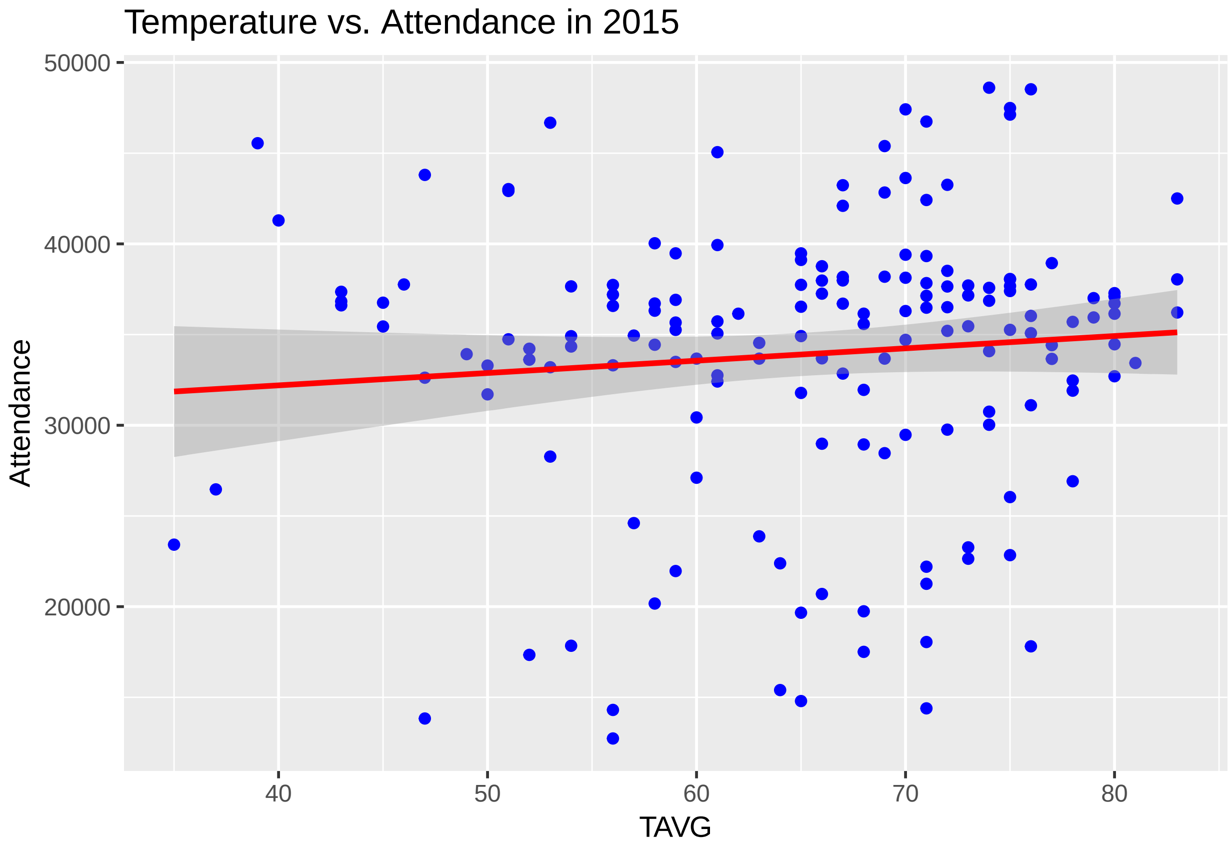
```
# 2017 Season
baseball_2017 <- baseball %>% filter(Year == 2017)
# Relationship between average temperature with attendance of season 2017
ggplot(baseball_2017, aes(TAVG, Attendance)) + geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red") + ggtitle("Temperature vs. Attendance in 2017")
```



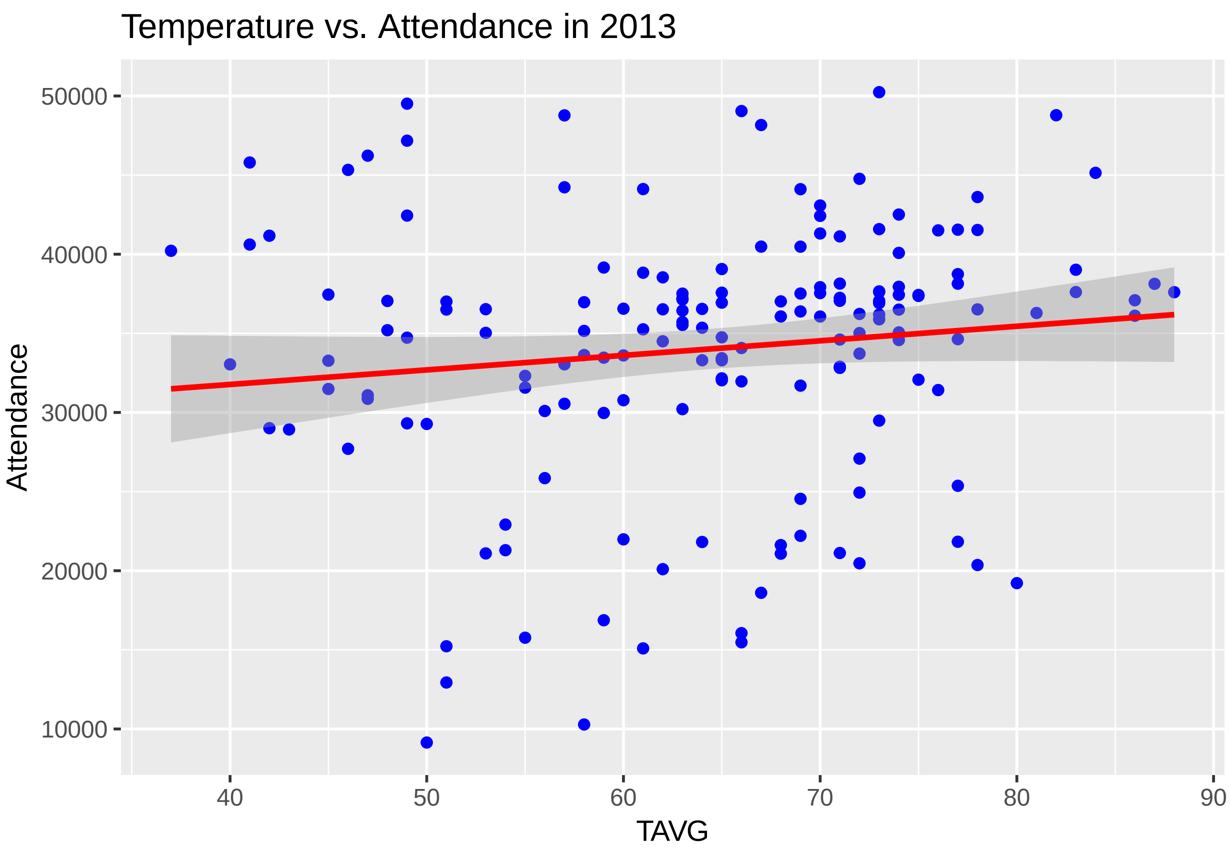
```
# 2016 Season
baseball_2016 <- baseball %>% filter(Year == 2016)
# Relationship between average temperature with attendance of season 2016
ggplot(baseball_2016, aes(TAVG, Attendance)) + geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red") + ggtitle("Temperature vs. Attendance in 2016")
```



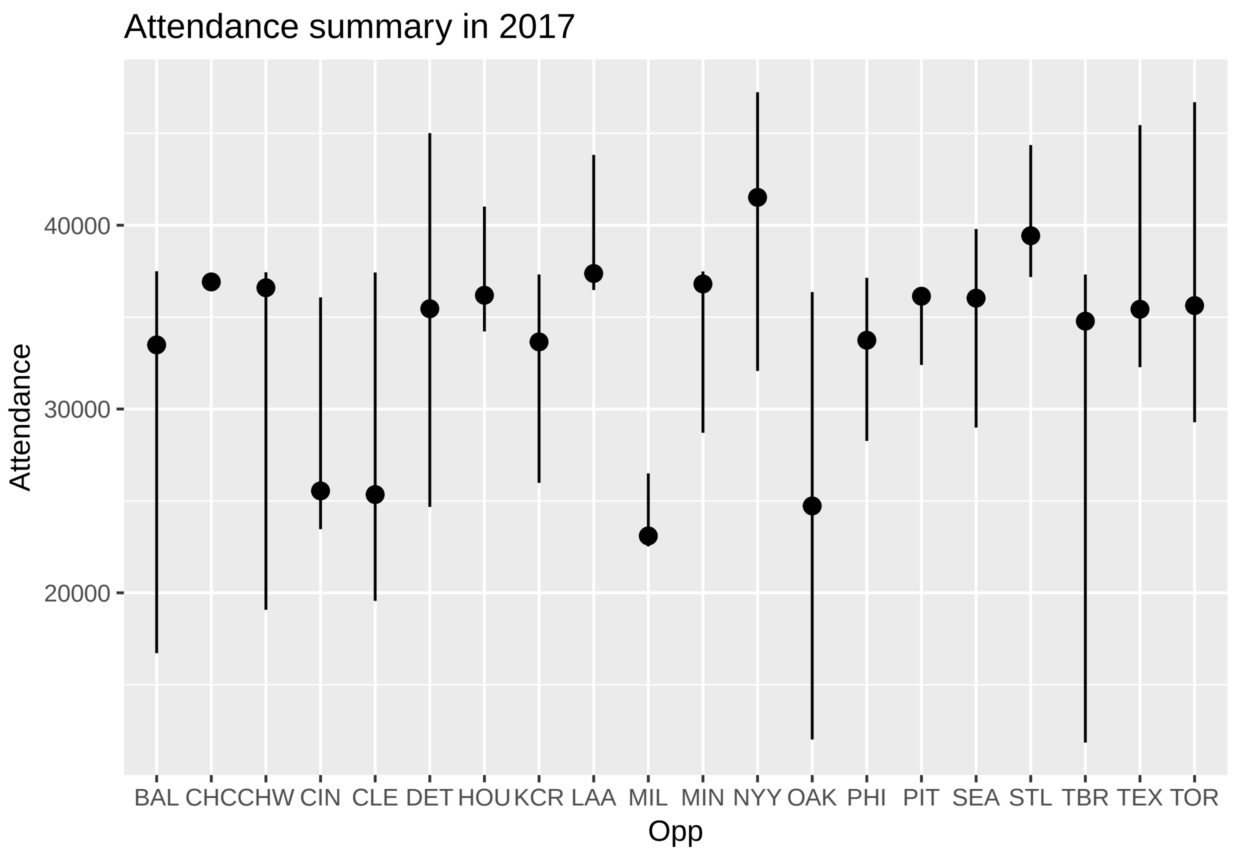
```
# 2015 Season
baseball_2015 <- baseball %>% filter(Year == 2015)
# Relationship between average temperature with attendance of season 2015
ggplot(baseball_2015, aes(TAVG, Attendance)) + geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red") + ggtitle("Temperature vs. Attendance in 2015")
```



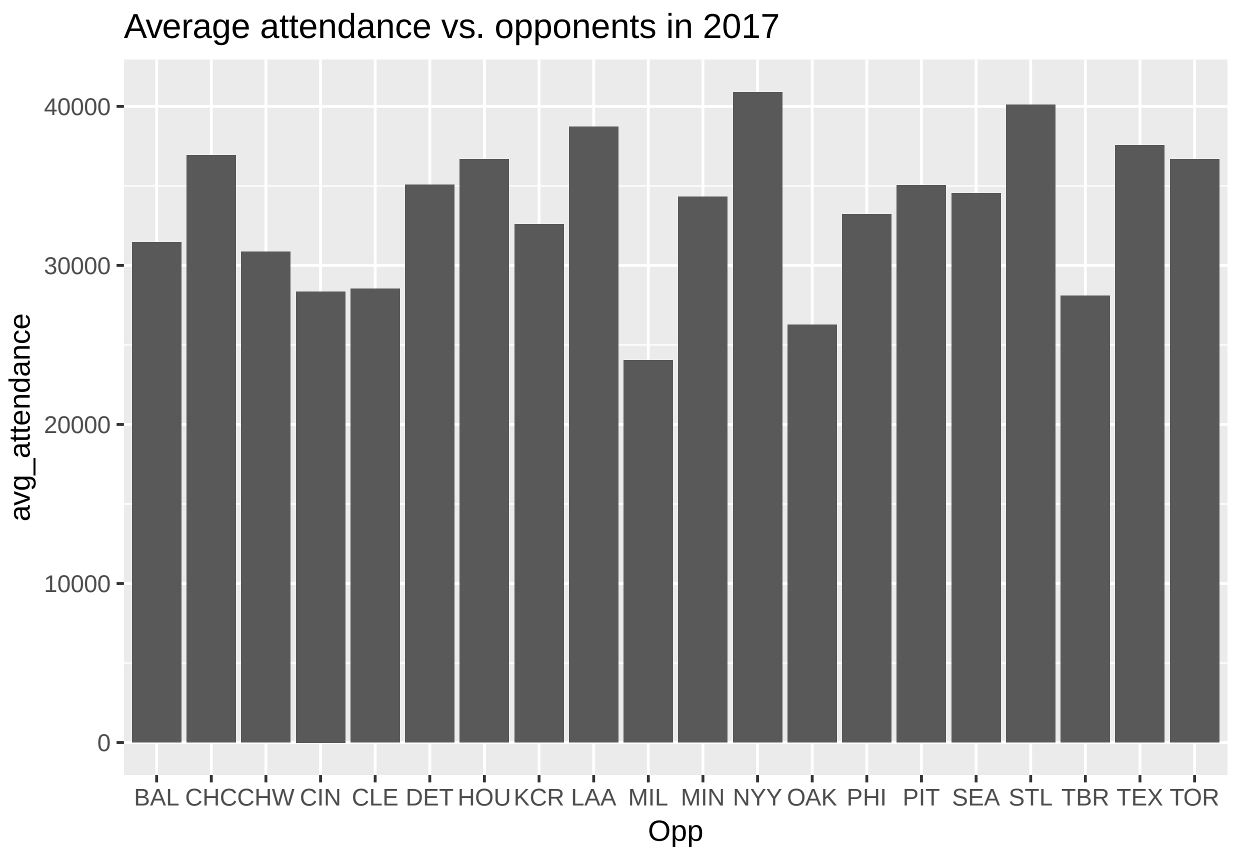
```
# 2013 Season
baseball_2013 <- baseball %>% filter(Year == 2013)
# Relationship between average temperature with attendance of season 2013
ggplot(baseball_2013, aes(TAVG, Attendance)) + geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red") + ggtitle("Temperature vs. Attendance in 2013")
```



```
# 2017
# Summary of attendance in seanson 2017 with different opponent.
ggplot(data = baseball_2017) +
  stat_summary(mapping = aes(x = Opp, y = Attendance), fun.ymin = min, fun.ymax = max, fun.y = median) +
  ggtitle("Attendance summary in 2017")
```



```
# Group by different opponents and arrange the attendance from high to low
baseball_opp <- baseball_2017 %>% group_by(Opp) %>% summarise(avg_attendance = mean(Attendance))
baseball_opp <- arrange(baseball_opp, desc(avg_attendance))

ggplot(baseball_opp, aes(Opp, avg_attendance)) +
  geom_bar(stat = "identity") +
  ggtitle("Average attendance vs. opponents in 2017")
```

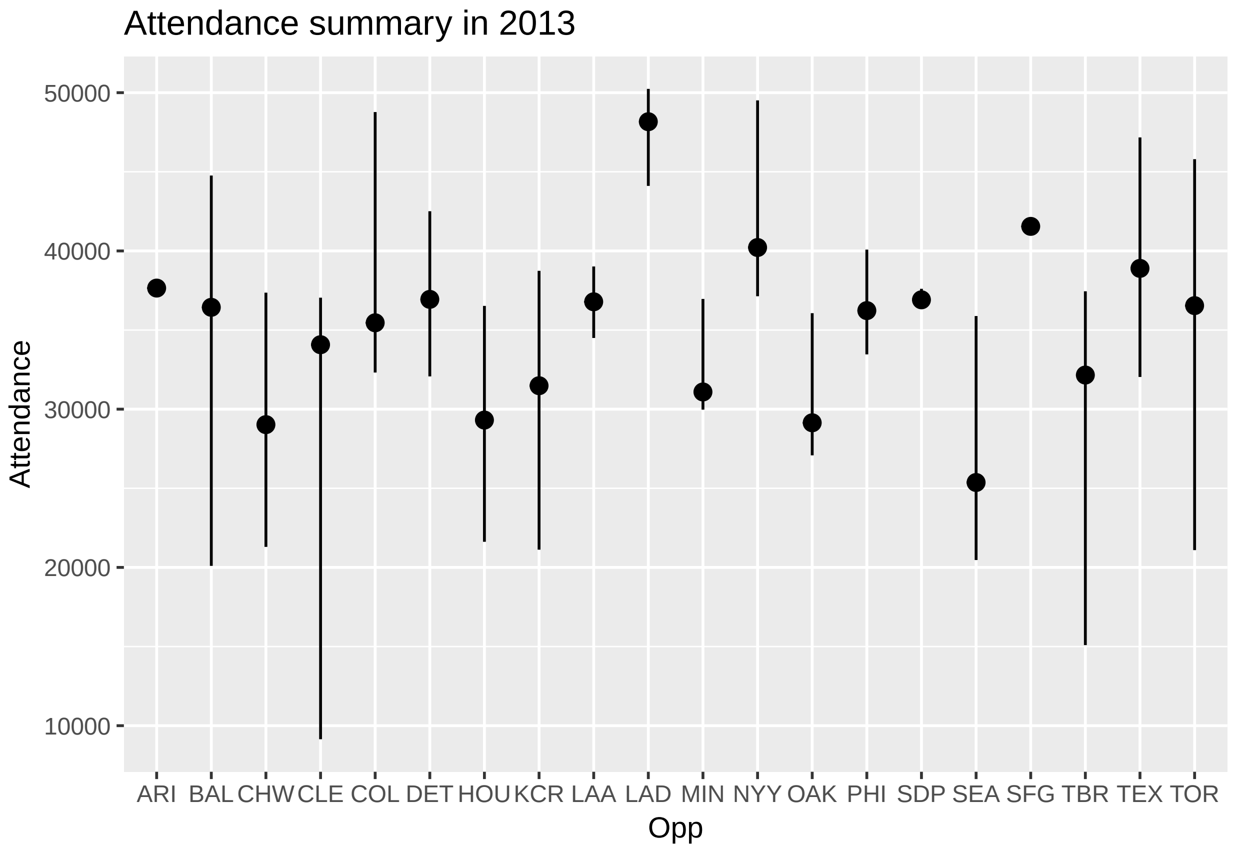


```r
kable(baseball_opp)
```

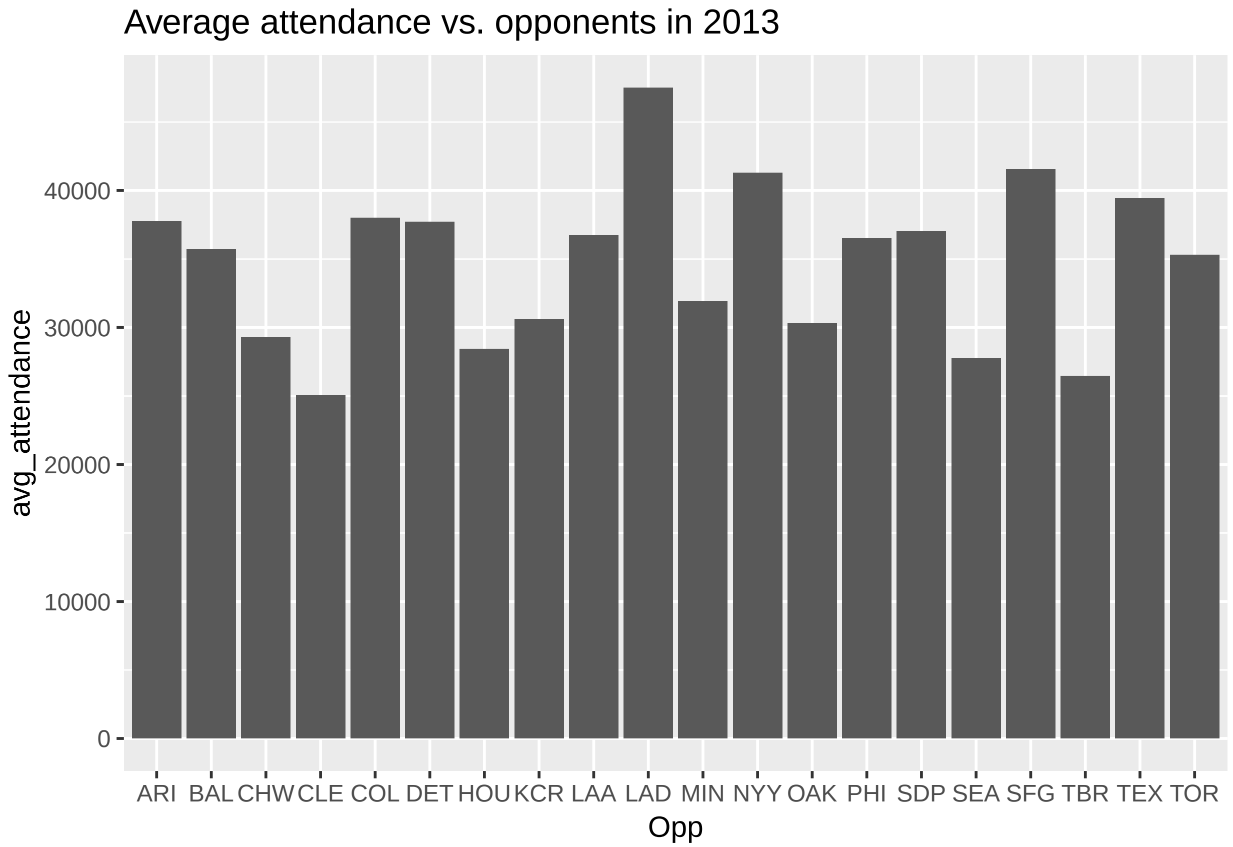
Opp	avg_attendance
NYY	40895.11
STL	40101.25
LAA	38731.67
TEX	37555.50
CHC	36915.33
TOR	36689.89
HOU	36668.43
DET	35063.00
PIT	35043.67
SEA	34548.33
MIN	34319.57
PHI	33222.50
KCR	32585.67
BAL	31444.68
CHW	30858.71
CLE	28529.29
CIN	28361.33
TBR	28101.89
OAK	26265.86
MIL	24039.33

```
# Based on the summary plot and average attendance table with different opponents,
```

```
# 2013
# Summary of attendance in seanson 2013 with different opponent.
ggplot(data = baseball_2013) +
  stat_summary(mapping = aes(x = Opp, y = Attendance), fun.ymin = min, fun.ymax = max, fun.y = median) +
  ggtitle("Attendance summary in 2013")
```



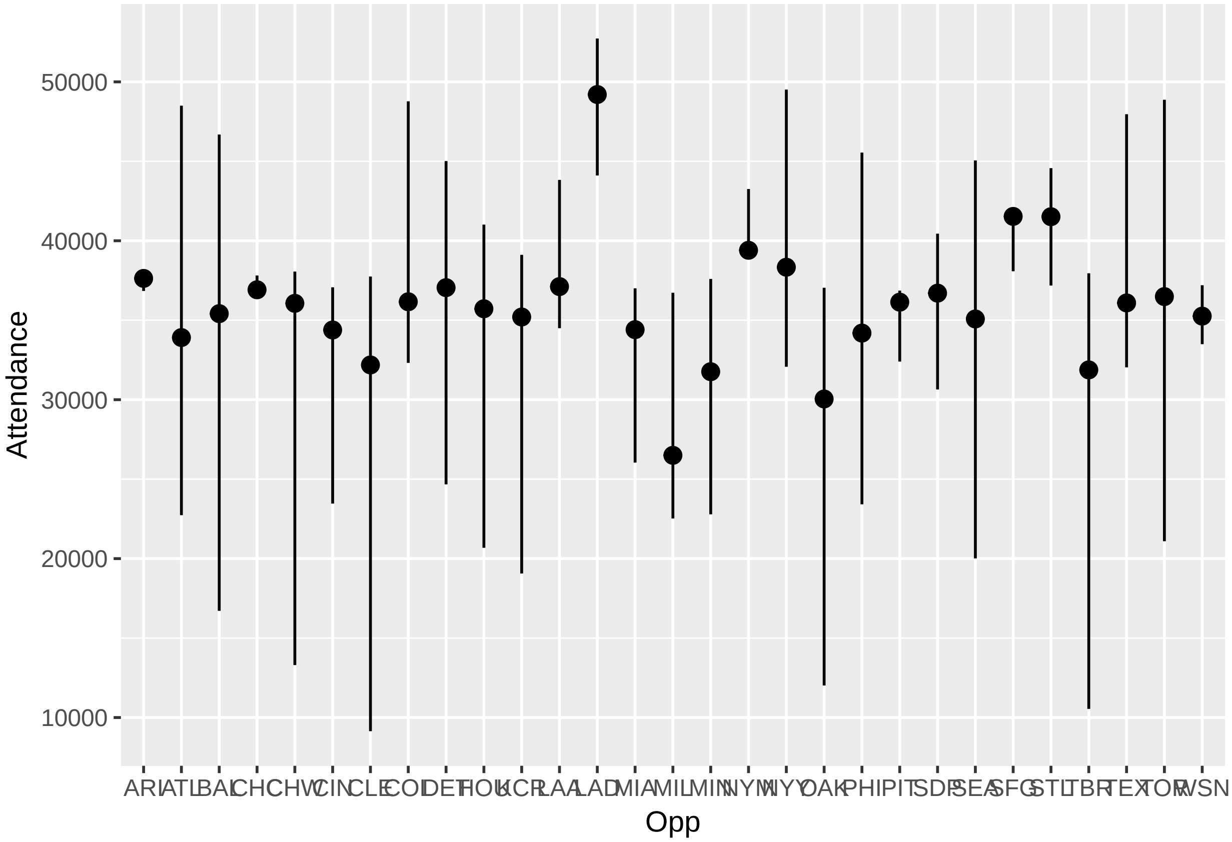
```
# Group by different opponents and arrange the attendance from high to low
baseball_opp <- baseball_2013 %>% group_by(Opp) %>% summarise(avg_attendance = mean(Attendance))
baseball_opp <- arrange(baseball_opp, desc(avg_attendance))

ggplot(baseball_opp, aes(Opp, avg_attendance)) +
  geom_bar(stat = "identity") +
  ggtitle("Average attendance vs. opponents in 2013")
```



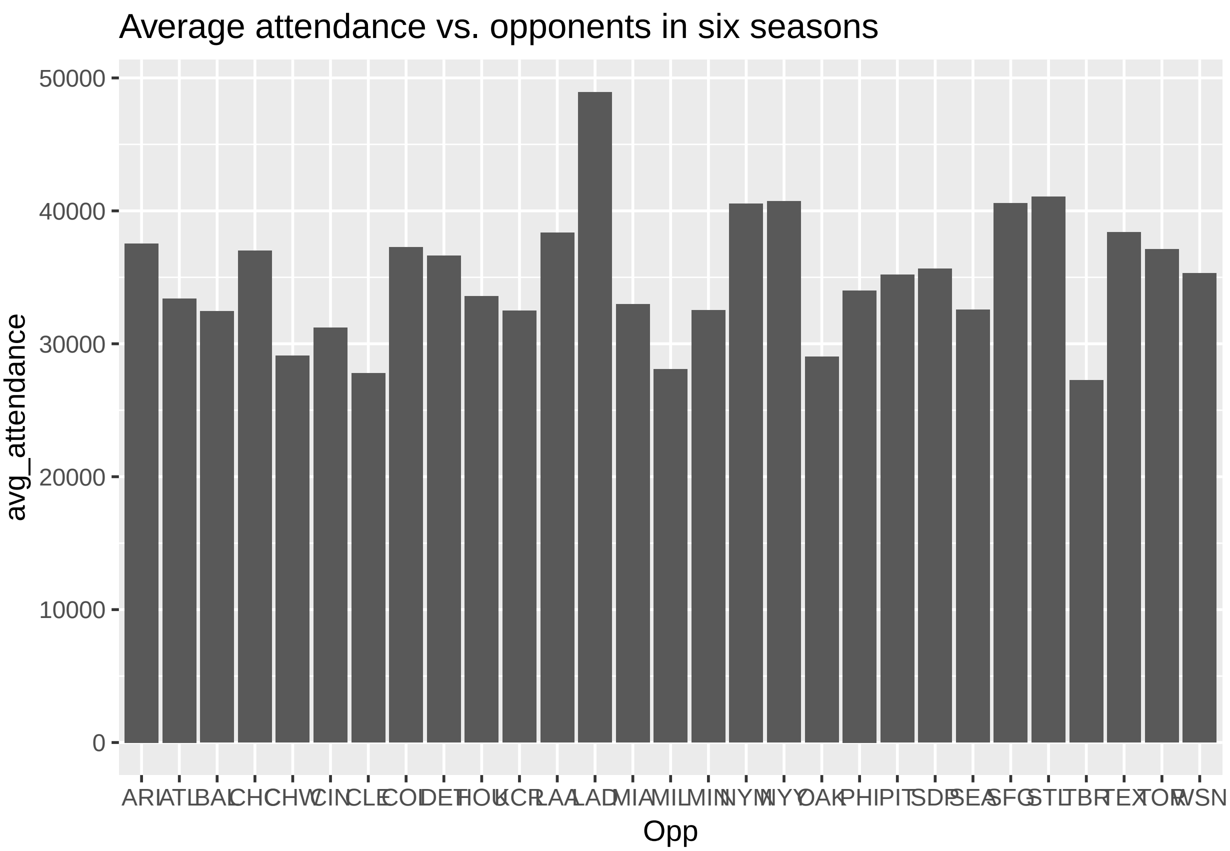
```r
kable(baseball_opp)
```

Opp	avg_attendance
LAD	47504.67
SFG	41556.00
NYY	41275.58
TEX	39420.83
COL	38002.00
ARI	37734.67
DET	37722.71
SDP	37005.33
LAA	36715.50
PHI	36501.00
BAL	35701.89
TOR	35295.84
MIN	31913.71
KCR	30598.29
OAK	30295.67
CHW	29262.83
HOU	28432.29
SEA	27752.00
TBR	26475.26
CLE	25034.14

```r
# Summary of attendance in 6 seansons with different opponent.
ggplot(data = baseball) +
  stat_summary(mapping = aes(x = Opp, y = Attendance), fun.ymin = min, fun.ymax = max, fun.y = median)
```



```r
# Group by different opponents and arrange the attendance from high to low
baseball_opp <- baseball %>% group_by(Opp) %>% summarise(avg_attendance = mean(Attendance))
baseball_opp <- arrange(baseball_opp, desc(avg_attendance))

ggplot(baseball_opp, aes(Opp, avg_attendance)) +
  geom_bar(stat = "identity") + ggtitle("Average attendance vs. opponents in six seasons")
```

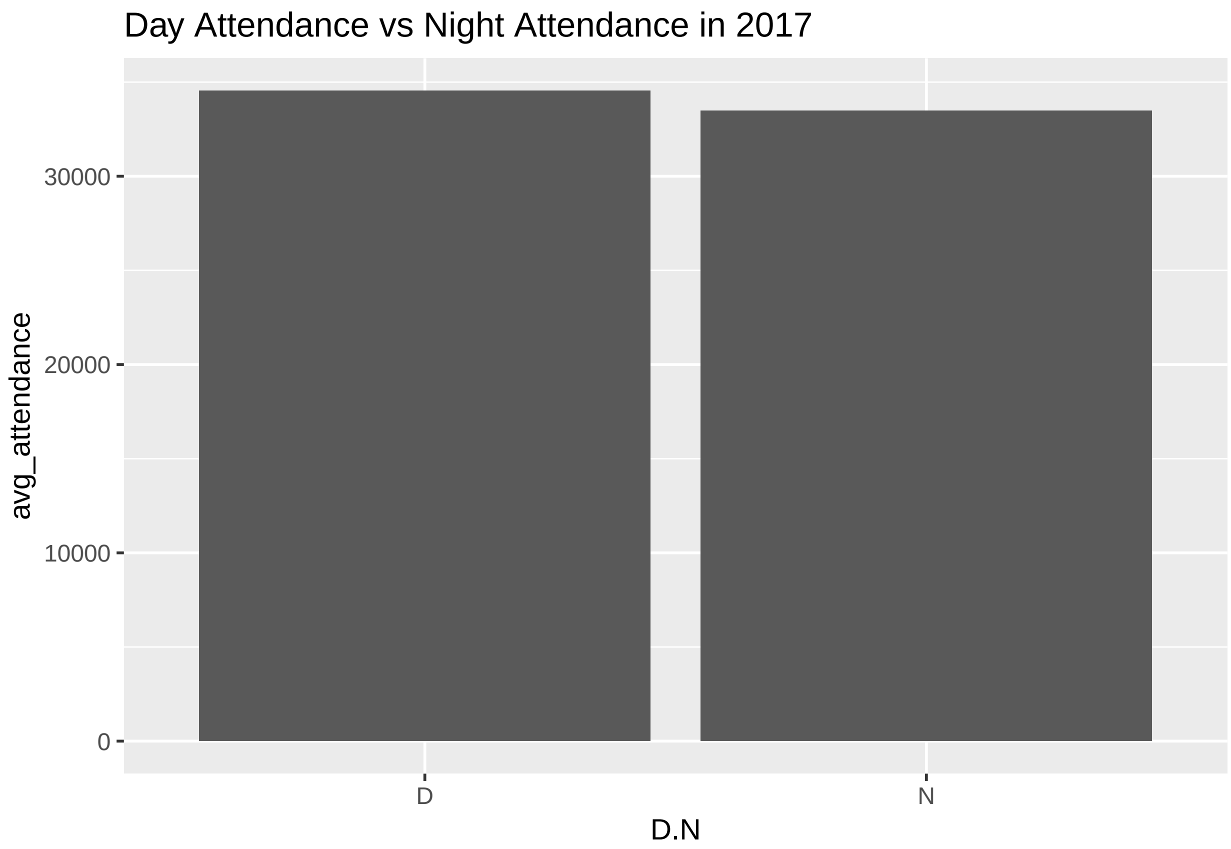


```
kable(baseball_opp)
```

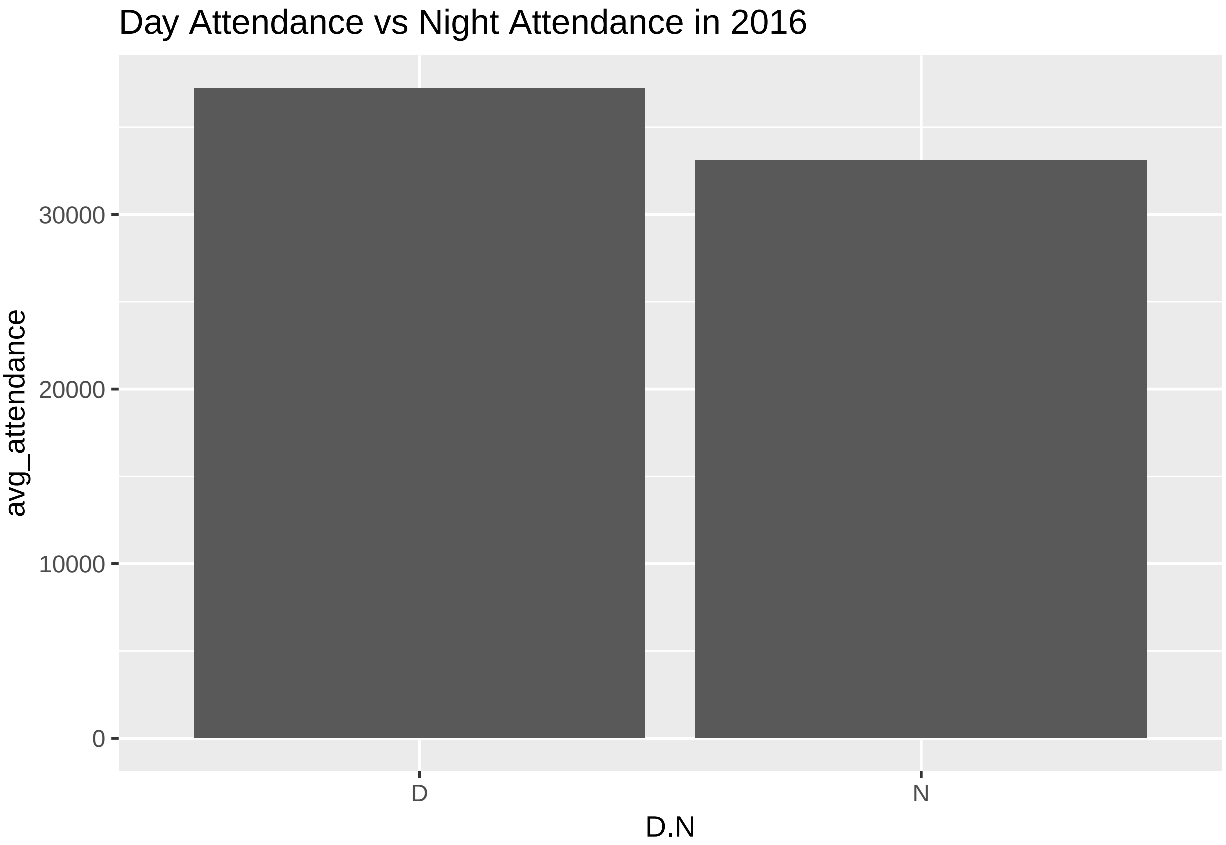
Opp	avg_attendance
LAD	48929.67
STL	41049.55
NYY	40730.42
SFG	40585.43
NYM	40531.33
TEX	38399.68
LAA	38360.64
ARI	37542.33
COL	37246.14
TOR	37120.76
CHC	37012.11
DET	36614.85
SDP	35628.00
WSN	35318.00
PIT	35178.56
PHI	34007.33
HOU	33573.24
ATL	33404.17
MIA	32966.00
SEA	32564.69
MIN	32530.34
KCR	32487.95
BAL	32460.79

Opp	avg_attendance
CIN	31201.70
CHW	29087.61
OAK	29011.79
MIL	28072.33
CLE	27771.47
TBR	27245.75

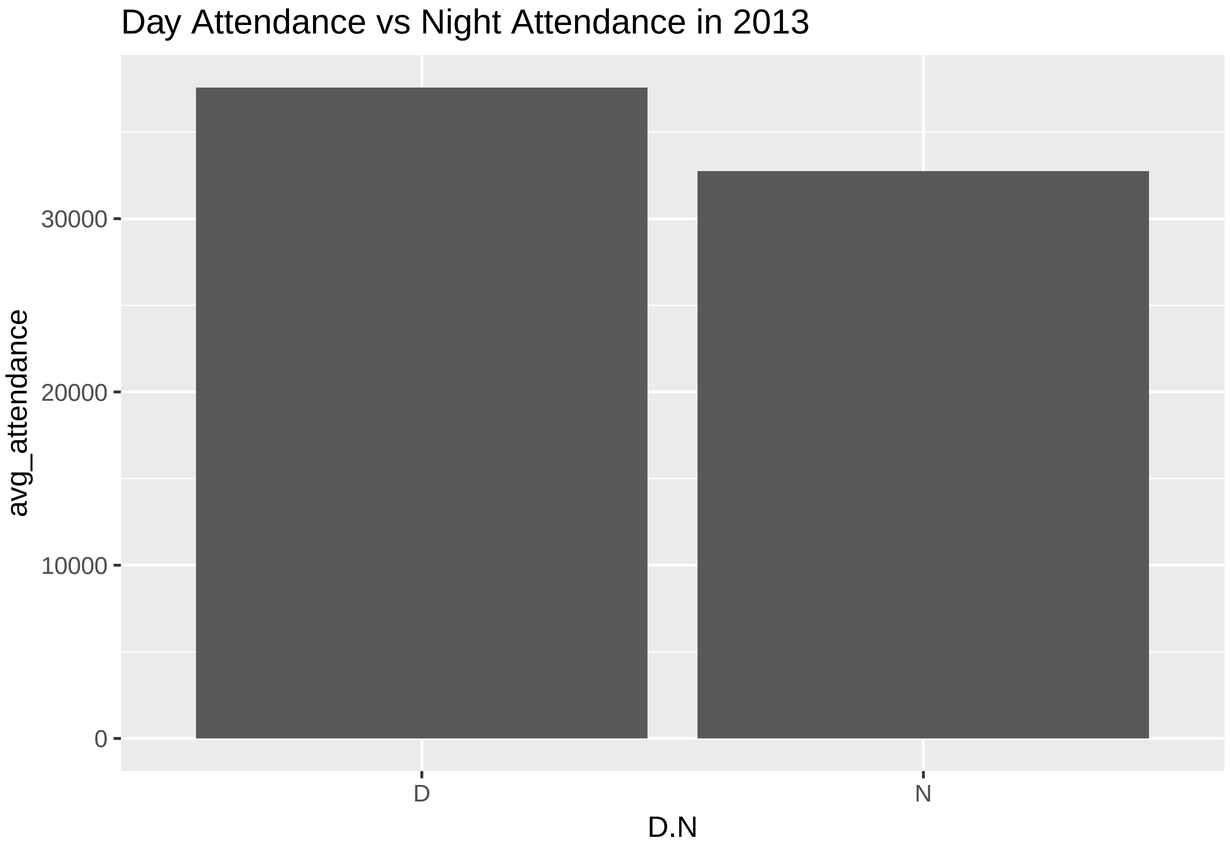
```
baseball_dn <- baseball_2017 %>% group_by(D.N) %>% summarise(avg_attendance = mean(Attendance))
baseball_dn <- arrange(baseball_dn, desc(avg_attendance))
ggplot(baseball_dn, aes(D.N, avg_attendance)) +
  geom_bar(stat = "identity") + ggtitle("Day Attendance vs Night Attendance in 2017")
```



```
baseball_dn <- baseball_2016 %>% group_by(D.N) %>% summarise(avg_attendance = mean(Attendance))
baseball_dn <- arrange(baseball_dn, desc(avg_attendance))
ggplot(baseball_dn, aes(D.N, avg_attendance)) +
  geom_bar(stat = "identity") + ggtitle("Day Attendance vs Night Attendance in 2016")
```



```
baseball_dn <- baseball_2013 %>% group_by(D.N) %>% summarise(avg_attendance = mean(Attendance))
baseball_dn <- arrange(baseball_dn, desc(avg_attendance))
ggplot(baseball_dn, aes(D.N, avg_attendance)) +
  geom_bar(stat = "identity")  + ggtitle("Day Attendance vs Night Attendance in 2013")
```


Day Attendance vs Night Attendance in 2013
avg_attendance
30000
20000
10000
0
D
N
D.N