Using Data to Improve MLB Attendance

Analya Ramirez

Step 1: Load the dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.2     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.4     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
dodgers<- read.csv("dodgers-2022.csv")
```

Step 2: Data exploration

```
str(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" ...
```

Step 3: Check for missing values

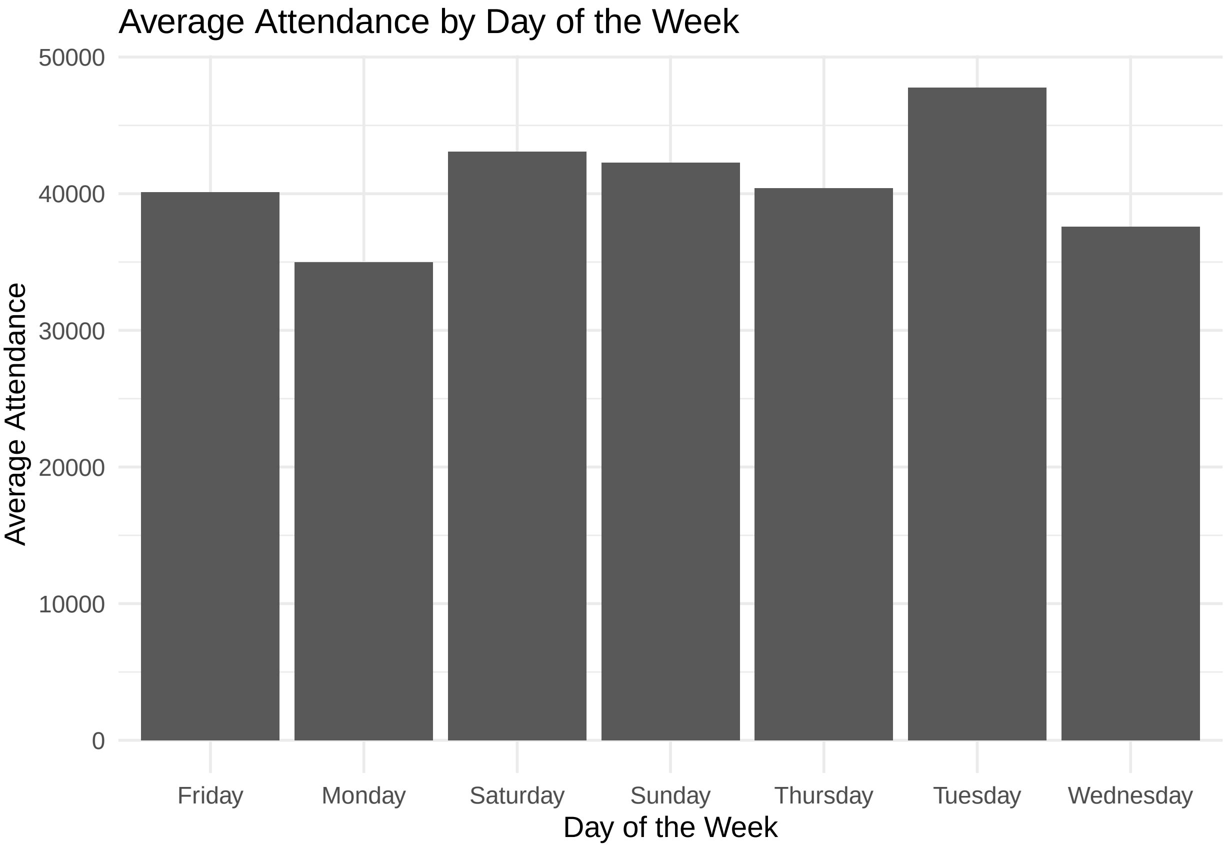
```
colSums(is.na(dodgers))
```

```
##       month         day      attend day_of_week    opponent        temp 
##           0           0           0           0           0           0 
##       skies   day_night         cap       shirt   fireworks  bobblehead 
##           0           0           0           0           0           0
```

There are no missing values in the dataset.

Step 4: Exploratory Data Analysis (EDA)

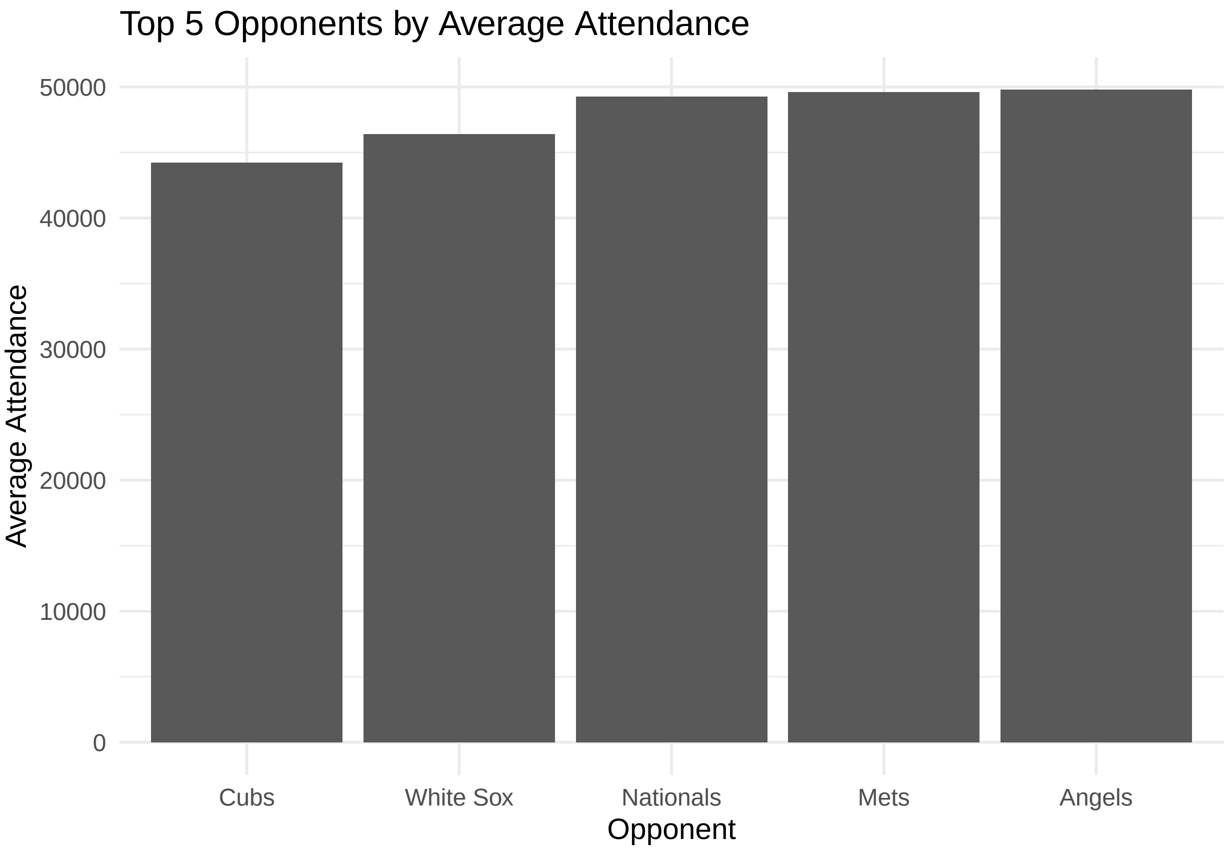
```
# Attendance by day of the week
dodgers %>%
  group_by(day_of_week) %>%
  summarise(average_attendance = mean(attend)) %>%
  ggplot(aes(x = day_of_week, y = average_attendance)) +
  geom_col() +
  labs(title = "Average Attendance by Day of the Week",
       x = "Day of the Week",
       y = "Average Attendance") +
  theme_minimal()
```



The average attendance was the highest for Tuesdays with an approximate average of 45,000 and the lowest average attendance was on Mondays.

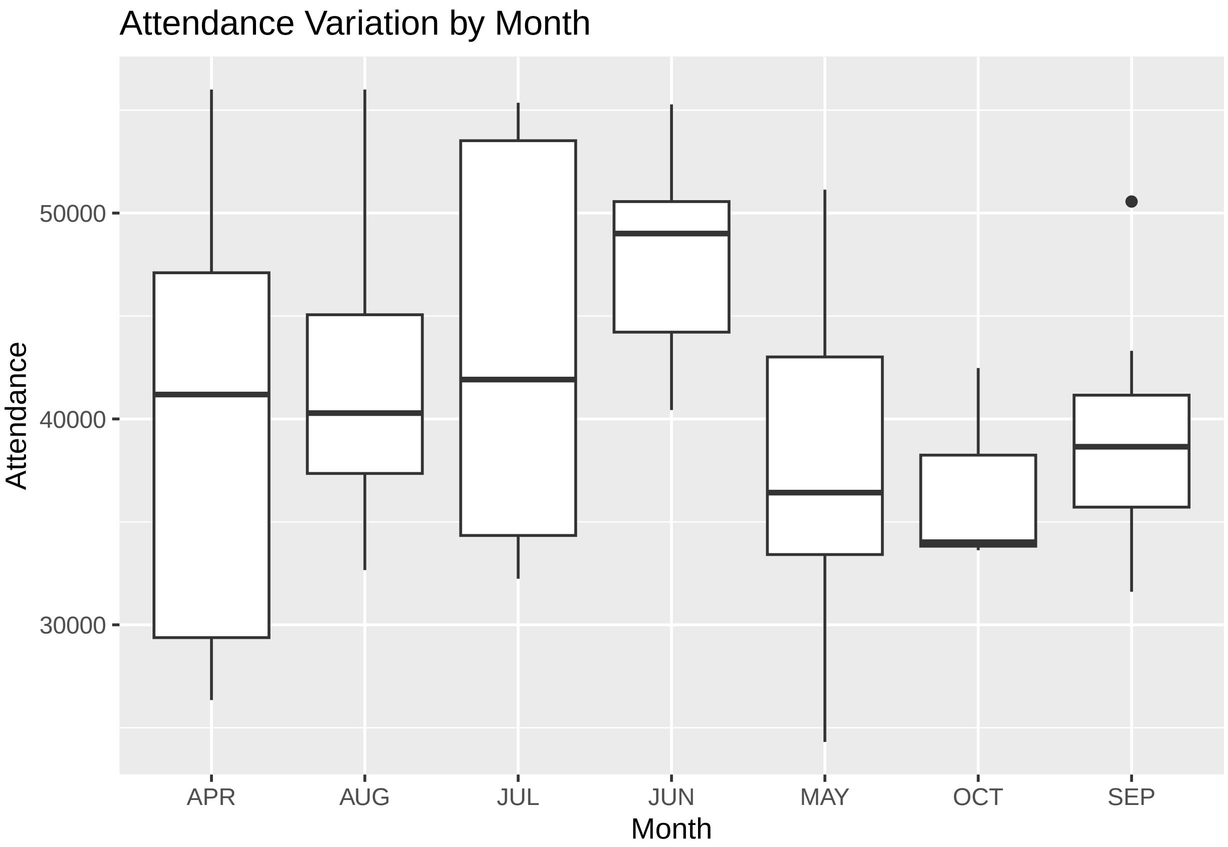
```
# Attendance by opponent
dodgers %>%
  group_by(opponent) %>%
  summarise(avg_attend = mean(attend)) %>%
  arrange(desc(avg_attend)) %>%
  top_n(5) %>%
  ggplot(aes(x = reorder(opponent, avg_attend), y = avg_attend)) +
  geom_col() +
  labs(title = "Top 5 Opponents by Average Attendance",
       x = "Opponent",
       y = "Average Attendance") +
  theme_minimal()
```

```
## Selecting by avg_attend
```



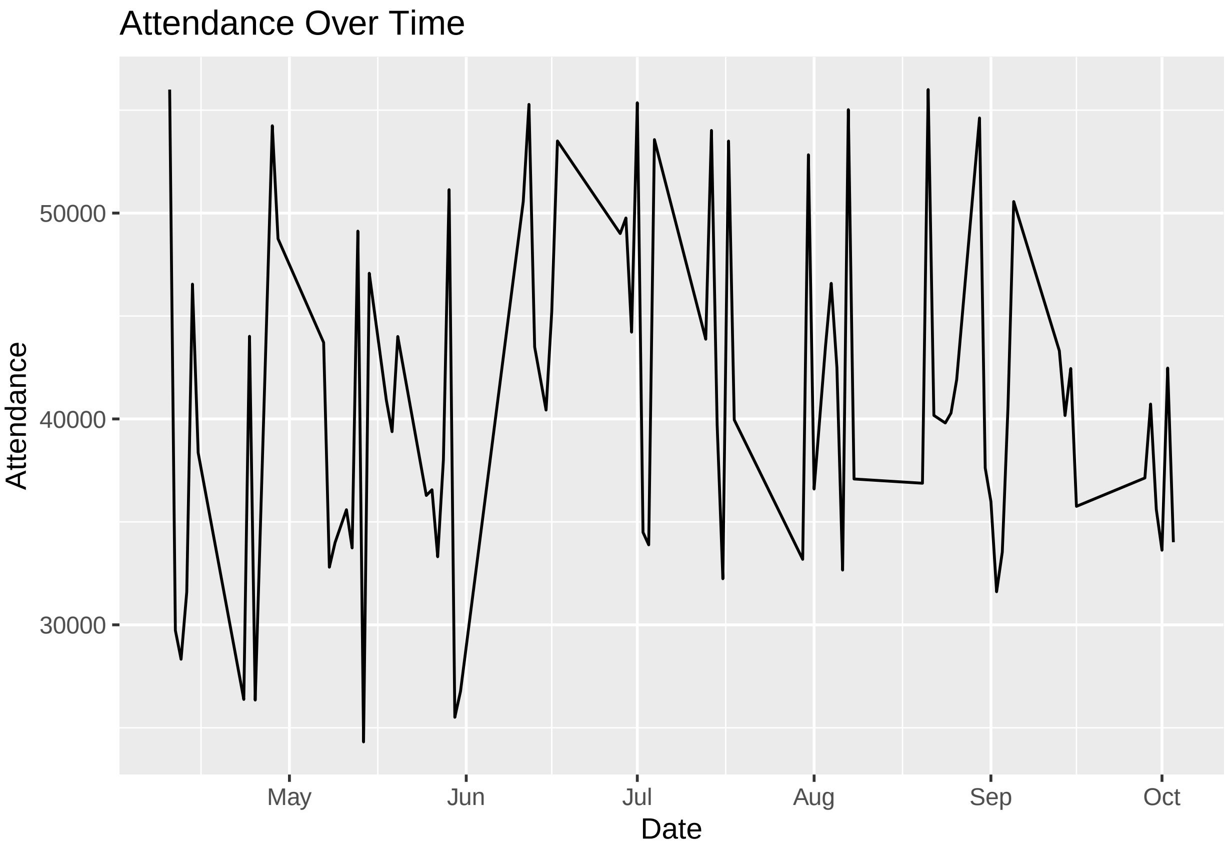
The top 5 opponents by average attendance are Angels, Mets, Nationals, White Sox and Cubs. Angels had the highest average attendance while the Cubs had the least.

```
# Attendance variation across different months
ggplot(dodgers, aes(x = month, y = attend)) +
  geom_boxplot() +
  labs(x = "Month", y = "Attendance", title = "Attendance Variation by Month")
```



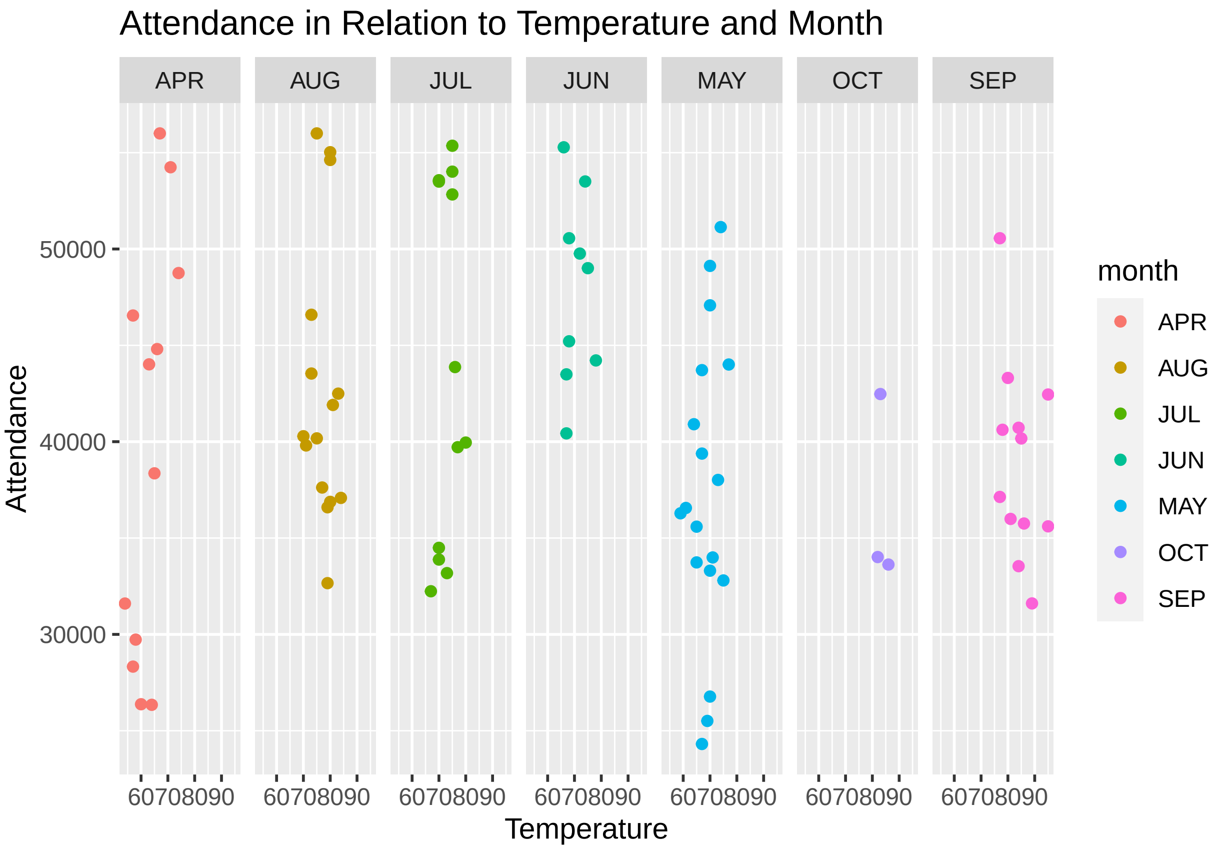
The average attendance is highest in June and the lowest in October.

```r
# Plot attendance over time
ggplot(dodgers, aes(x = as.Date(paste(month, day), format = "%b %d"), y = attend)) +
  geom_line() +
  labs(x = "Date", y = "Attendance", title = "Attendance Over Time")
```



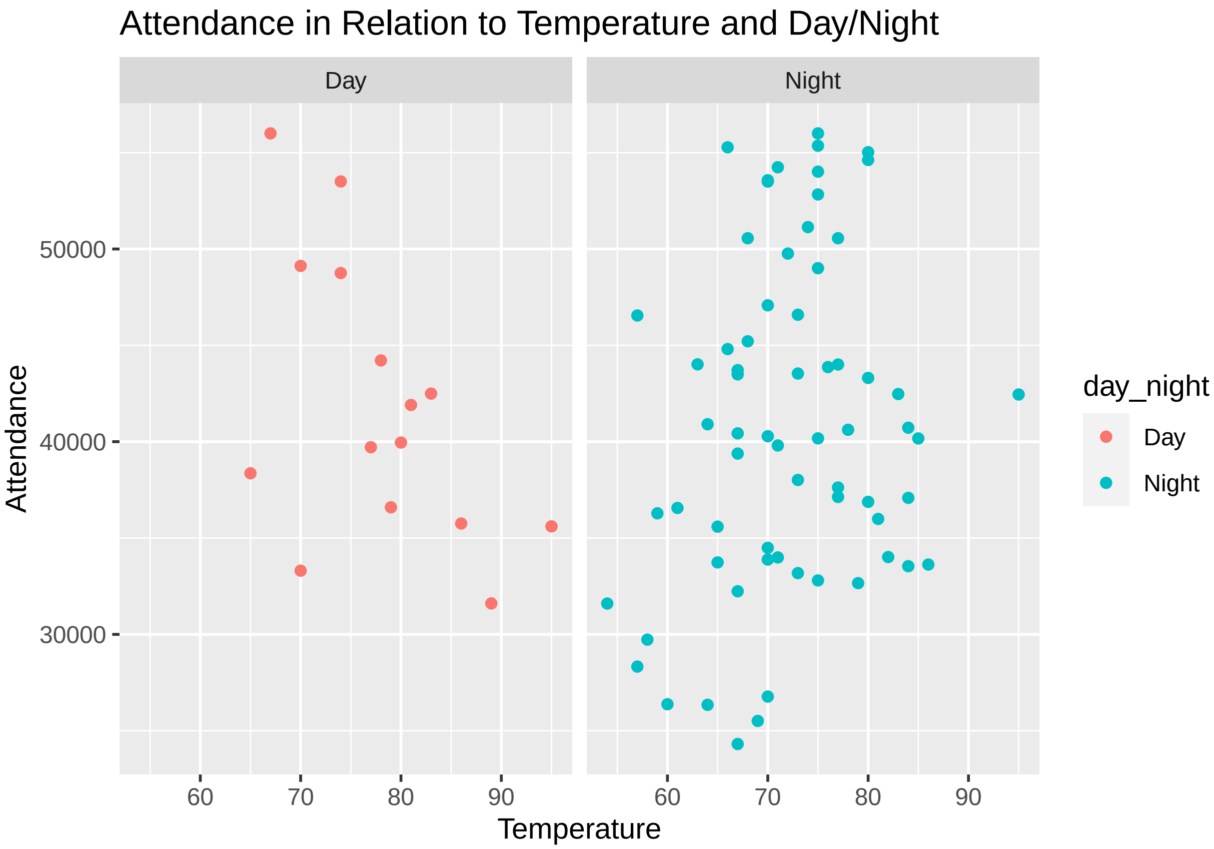
The highest attendance is showing a decreasing trend overtime.

```
## Attendance in relation to temperature and month
dodgers$month<-as.factor(dodgers$month)
ggplot(dodgers, aes(x=temp, y=attend, color=month)) + geom_point() + facet_grid(~month)+
  labs(x = "Temperature", y = "Attendance", title = "Attendance in Relation to Temperature and Month")
```



October had the least attendance and May had the most attendance in relation to temperature. The relationship between temperature and attendance in relation to months is non linear.

```
## Attendance in relation to temperature and day_night
dodgers$day_night<-as.factor(dodgers$day_night)
ggplot(dodgers, aes(x=temp, y=attend, color=day_night)) + geom_point() + facet_grid(~day_night)+
  labs(x = "Temperature", y = "Attendance", title = "Attendance in Relation to Temperature and Day/Nigh
```



The relationship between temperature and attendance in relation to day and night is linear.

Theres a negative linear relationship between temperature and attendance during the day, meaning increase in temperature causes decrease in attendance.

Theres a positive linear relationship between temperature and attendance during the night, meaning increase in temperature causes increase in attendance.

```
# Correlation Analysis
# Calculate correlations
corr1 <- cor(dodgers[, c("attend", "temp")])
corr1
```

```
##             attend       temp
## attend 1.00000000 0.09895073
## temp   0.09895073 1.00000000
```

There is a high positive correlation (0.098) between temperature and attendance, meaning increase in temperature causes a increase in attendance.

```
# Calculate correlations
corr2 <- cor(dodgers[, c("attend", "day")])
corr2
```

```
##             attend        day
## attend 1.00000000 0.02709298
## day    0.02709298 1.00000000
```

There is a low positive correlation (0.027) between attendance and day, meaning increase in day causes a slight increase in attendance.

Step 7: Reccomendations

Recommendation 1: Opponent Selection

Schedule more games against popular opponents that is Angels, Mets, Nationals, White Sox and Cubs, as they tend to attract higher attendance.

Recommendation 2: Game Scheduling

Schedule games on Tuesdays and the month of June since historically they have higher attendance.

Recommendation 3: Game scheduling in relation to temperature

Schedule games in May since increase in temperature means increase in attendance for the month of May.

Recommendation 4: Game schedule in relation to day_night and temperature

Schedule games at night since increase temperature at night causes increase in attendance compared to during the day.

Recommendation 5: Enhancing Fan Experience

Improve the overall fan experience by investing in stadium facilities, providing engaging entertainment during games, and creating a welcoming and inclusive atmosphere for all fans.