In [441]:
```
#Paul Galvez
#Class: DSC 630-T301
#Date: 6/22/23
#Week 3 Assingment 3.2
```

In [442]:
```
#In this assignment, you will be using data on the Los Angeles Dodgers Major League Baseball (MLB) team locate
#dodgers.csv. Use this data to make a recommendation to management on how to improve attendance.
#Tell a story with your analysis and clearly explain the steps you take to arrive at your conclusion.
#This is an open-ended question, and there is no one right answer. You are welcome to do additional
#research and/or use domain knowledge to assist your analysis, but clearly state any assumptions you make.
#You can use R or Python to complete this assignment. Submit your code and output to the submission link.
#Make sure to add comments to all your code and to document your steps, process, and analysis.
```

In [443]:
```
#loading the libraries that are needed for the analysis

import matplotlib as mpl
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
import pandas as pd
from cycler import cycler
```

In [444]:
```
#Loading the CSV data 'dodgers.csv'

df = pd.read_csv('dodgers.csv')
df
```

Out[444]:

	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks	bobblehead
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO	NO
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	NO	NO	NO	NO
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	NO	NO	NO	NO
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	NO	NO	YES	NO
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	NO	NO	NO	NO
...	...	...	...	...	...	...	...	...	...	...	...	...
76	SEP	29	40724	Saturday	Rockies	84	Cloudy	Night	NO	NO	NO	NO
77	SEP	30	35607	Sunday	Rockies	95	Clear	Day	NO	NO	NO	NO
78	OCT	1	33624	Monday	Giants	86	Clear	Night	NO	NO	NO	NO
79	OCT	2	42473	Tuesday	Giants	83	Clear	Night	NO	NO	NO	NO
80	OCT	3	34014	Wednesday	Giants	82	Cloudy	Night	NO	NO	NO	NO

81 rows × 12 columns

In [446]:
```
#making sure the data has been loaded properly

df.head(15)
```

Out[446]:

	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks	bobblehead
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO	NO
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	NO	NO	NO	NO
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	NO	NO	NO	NO
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	NO	NO	YES	NO
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	NO	NO	NO	NO
5	APR	15	38359	Sunday	Padres	65	Clear	Day	NO	NO	NO	NO
6	APR	23	26376	Monday	Braves	60	Cloudy	Night	NO	NO	NO	NO
7	APR	24	44014	Tuesday	Braves	63	Cloudy	Night	NO	NO	NO	NO
8	APR	25	26345	Wednesday	Braves	64	Cloudy	Night	NO	NO	NO	NO
9	APR	27	44807	Friday	Nationals	66	Clear	Night	NO	NO	YES	NO
10	APR	28	54242	Saturday	Nationals	71	Clear	Night	NO	NO	NO	YES
11	APR	29	48753	Sunday	Nationals	74	Clear	Day	NO	YES	NO	NO
12	MAY	7	43713	Monday	Giants	67	Clear	Night	NO	NO	NO	NO
13	MAY	8	32799	Tuesday	Giants	75	Clear	Night	NO	NO	NO	NO
14	MAY	9	33993	Wednesday	Giants	71	Clear	Night	NO	NO	NO	NO

In [447]:
```
#possible variables that could impact attendance include
#date, opponent, day of the week, weather conditions, promotional events,
#team performance, and ticket prices.

df.columns
```

Out[447]:
```
Index(['month', 'day', 'attend', 'day_of_week', 'opponent', 'temp', 'skies',
       'day_night', 'cap', 'shirt', 'fireworks', 'bobblehead'],
      dtype='object')
```

In [448]:
```
df.describe()
```

Out[448]:

	day	attend	temp
count	81.000000	81.000000	81.000000
mean	16.135802	41040.074074	73.148148
std	9.605666	8297.539460	8.317318
min	1.000000	24312.000000	54.000000
25%	8.000000	34493.000000	67.000000
50%	15.000000	40284.000000	73.000000
75%	25.000000	46588.000000	79.000000
max	31.000000	56000.000000	95.000000

In [449]:
```
df.shape
```

Out[449]: (81, 12)

In [450]:
```
#renamed day_of_week to day of week and day_night to day or night for readability.

df.rename(columns={'day_of_week': 'day of week', 'day_night': 'day or night'}, inplace=True)
df_two = df.head(20)
df_two
```

Out[450]:

	month	day	attend	day of week	opponent	temp	skies	day or night	cap	shirt	fireworks	bobblehead
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO	NO
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	NO	NO	NO	NO
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	NO	NO	NO	NO
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	NO	NO	YES	NO
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	NO	NO	NO	NO
5	APR	15	38359	Sunday	Padres	65	Clear	Day	NO	NO	NO	NO
6	APR	23	26376	Monday	Braves	60	Cloudy	Night	NO	NO	NO	NO
7	APR	24	44014	Tuesday	Braves	63	Cloudy	Night	NO	NO	NO	NO
8	APR	25	26345	Wednesday	Braves	64	Cloudy	Night	NO	NO	NO	NO
9	APR	27	44807	Friday	Nationals	66	Clear	Night	NO	NO	YES	NO
10	APR	28	54242	Saturday	Nationals	71	Clear	Night	NO	NO	NO	YES
11	APR	29	48753	Sunday	Nationals	74	Clear	Day	NO	YES	NO	NO
12	MAY	7	43713	Monday	Giants	67	Clear	Night	NO	NO	NO	NO
13	MAY	8	32799	Tuesday	Giants	75	Clear	Night	NO	NO	NO	NO
14	MAY	9	33993	Wednesday	Giants	71	Clear	Night	NO	NO	NO	NO
15	MAY	11	35591	Friday	Rockies	65	Clear	Night	NO	NO	YES	NO
16	MAY	12	33735	Saturday	Rockies	65	Clear	Night	NO	NO	NO	NO
17	MAY	13	49124	Sunday	Rockies	70	Clear	Day	NO	NO	NO	NO
18	MAY	14	24312	Monday	Snakes	67	Clear	Night	NO	NO	NO	NO
19	MAY	15	47077	Tuesday	Snakes	70	Clear	Night	NO	NO	NO	YES

In [452]:
```
#dropping duplicates from the data.

df = df.drop_duplicates()
df
```

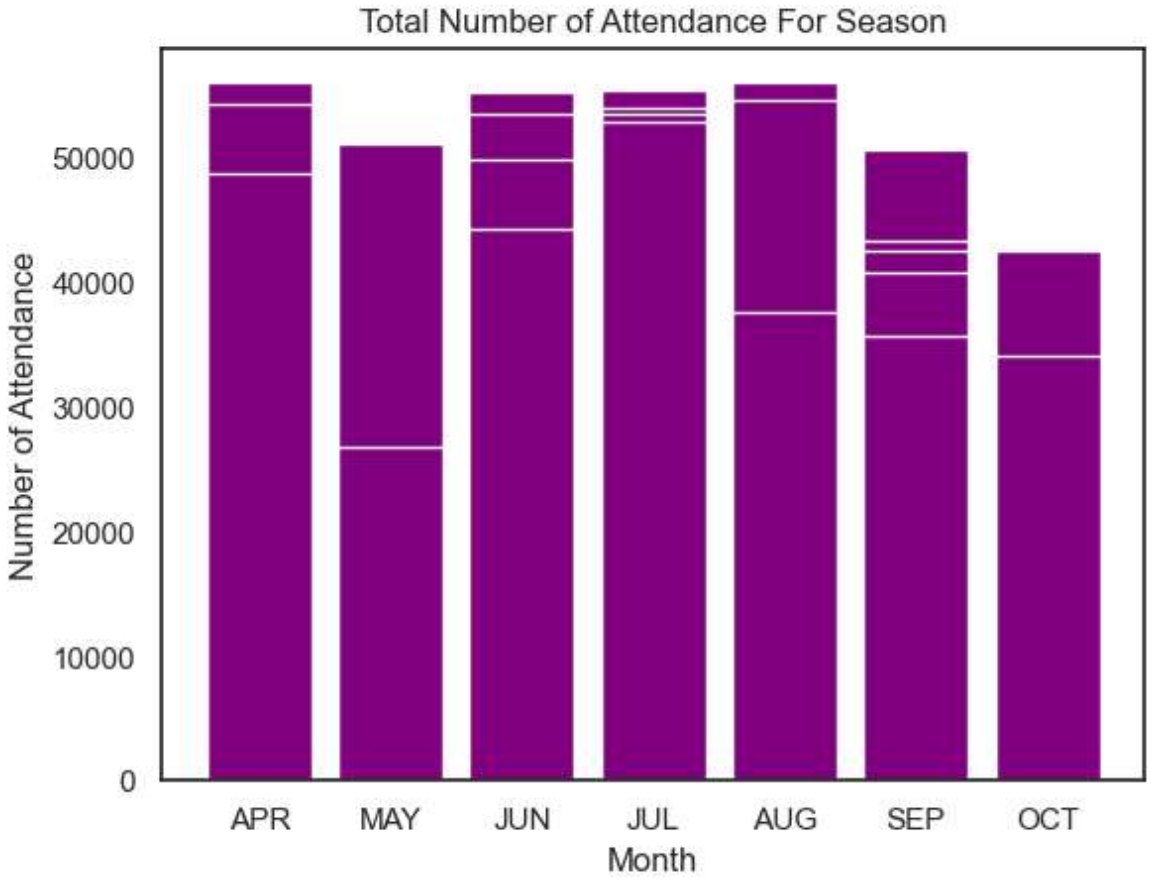
Out[452]:

	month	day	attend	day of week	opponent	temp	skies	day or night	cap	shirt	fireworks	bobblehead
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO	NO
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	NO	NO	NO	NO
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	NO	NO	NO	NO
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	NO	NO	YES	NO
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	NO	NO	NO	NO
...	...	...	...	...	...	...	...	...	...	...	...	...
76	SEP	29	40724	Saturday	Rockies	84	Cloudy	Night	NO	NO	NO	NO
77	SEP	30	35607	Sunday	Rockies	95	Clear	Day	NO	NO	NO	NO
78	OCT	1	33624	Monday	Giants	86	Clear	Night	NO	NO	NO	NO
79	OCT	2	42473	Tuesday	Giants	83	Clear	Night	NO	NO	NO	NO
80	OCT	3	34014	Wednesday	Giants	82	Cloudy	Night	NO	NO	NO	NO

81 rows × 12 columns

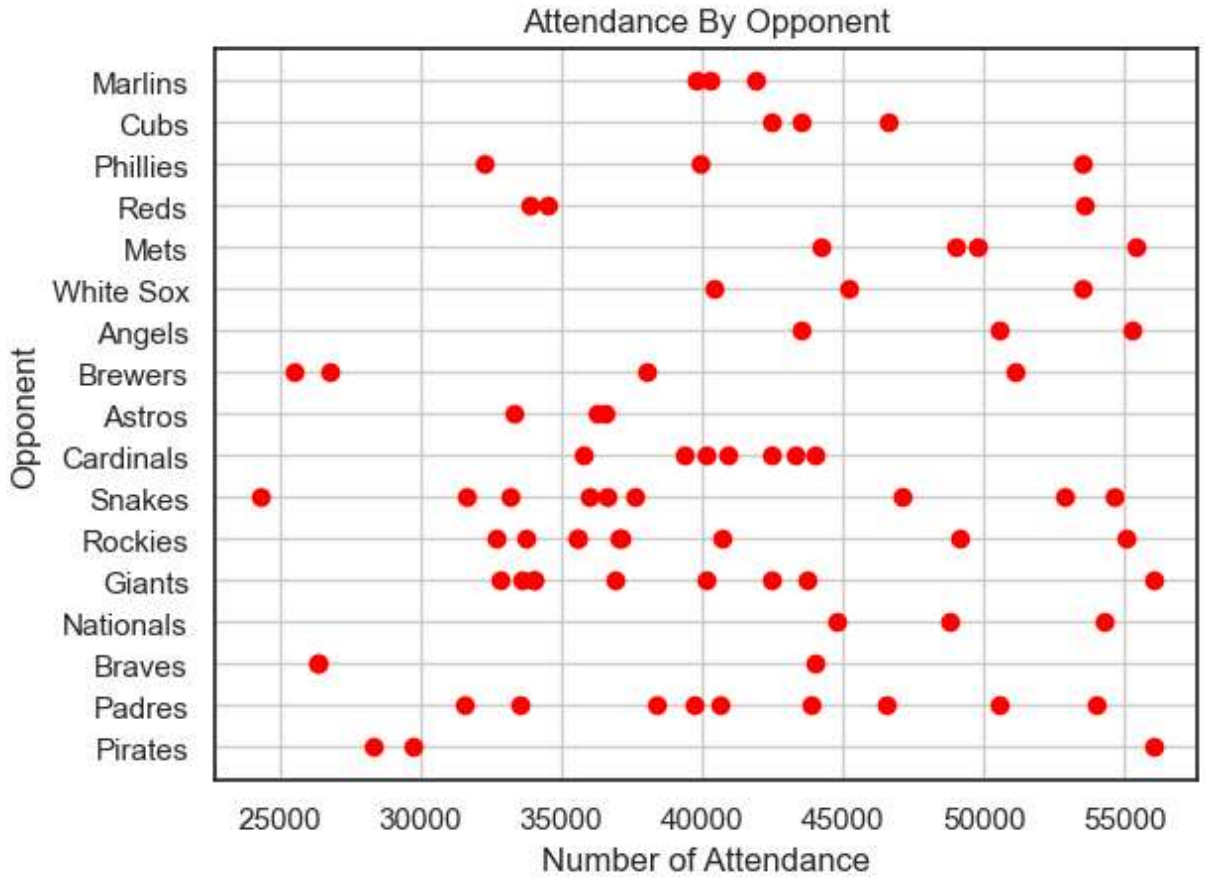
In [453]:
```
#a look over the course of the season for overall attendance from April to October. We can see
#attendance stays higher during the summer months while attendance comes down in the fall months
#September and October.

plt.bar(df['month'], df['attend'], color='purple')
plt.xlabel('Month')
plt.ylabel('Number of Attendance')
plt.title('Total Number of Attendance For Season')
plt.show()
```



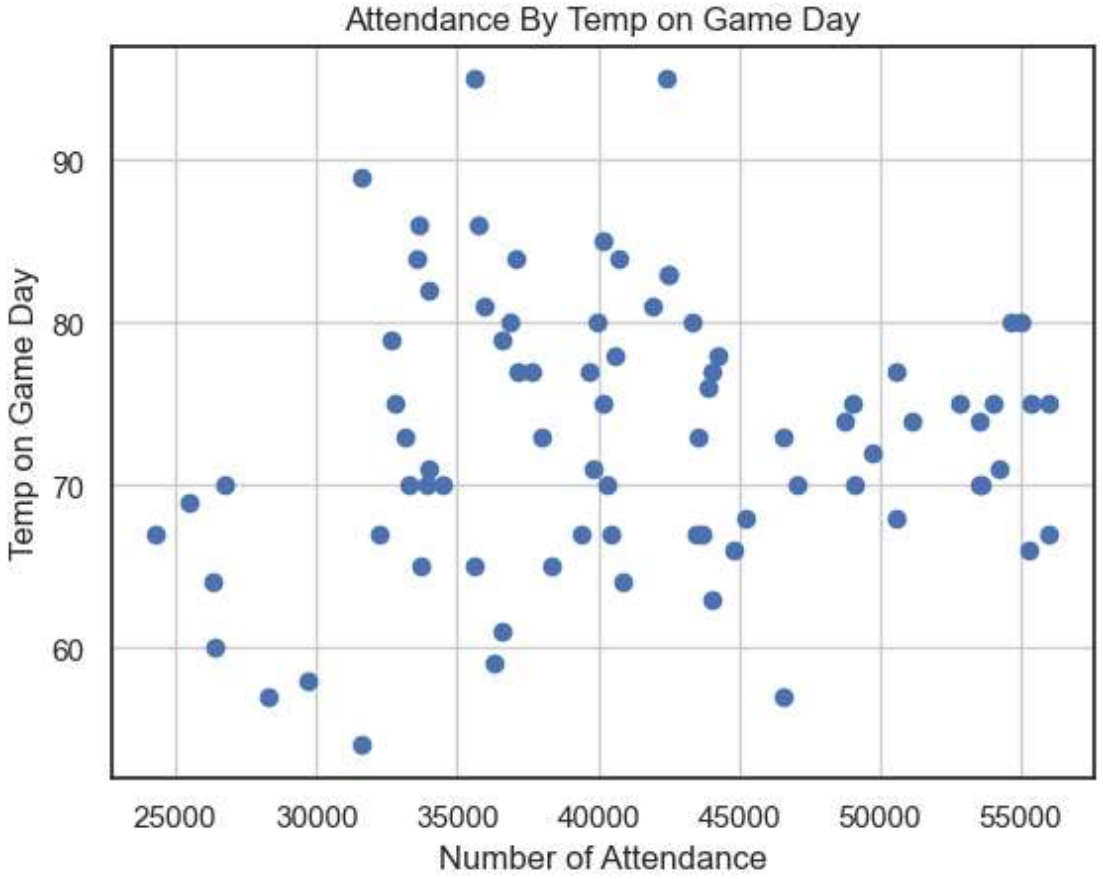
In [455]:
```
#scatter plot showing attendance by opponent. We can see attendance is higher when the Dodgers are playing tea
#well established franchise history in the leauge. Opponents such as the Pirates, Reds, Giants, and Padres hav
#highest numbers of attendance during the season. On the other hand, teams like the snakes and Brewers have mu
#numbers for attendance.

plt.scatter(df['attend'], df['opponent'], color='red')
plt.xlabel('Number of Attendance')
plt.ylabel('Opponent')
plt.title('Attendance By Opponent')
plt.grid()
plt.show()
```



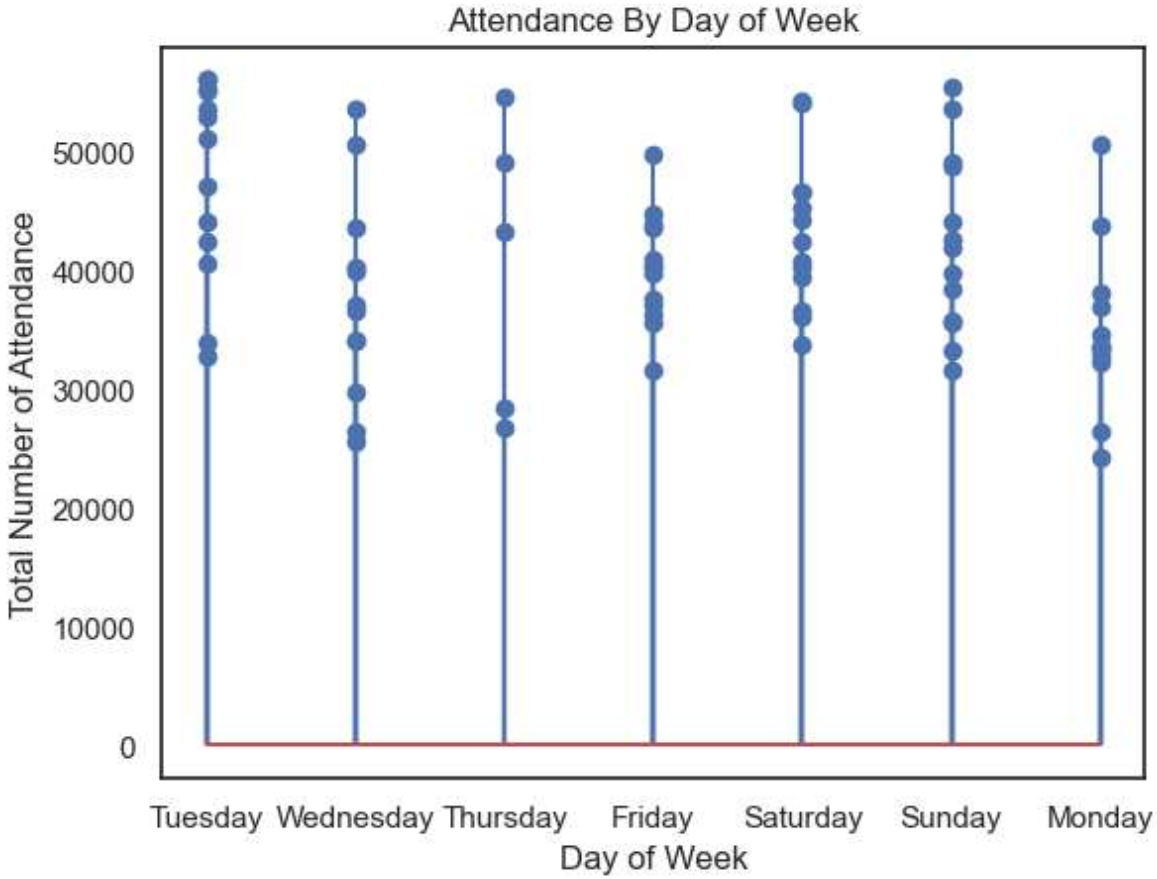
In [456]:
```
#by comparison the scatter plot below shows attendance by temp. on game day. We wee most of the data points ar
#range of 60 - 90 degrees showing most people attended games throughout the season when the temp on game day l
#those two temps. There are some outliers on either end of the temp scale. But those games were not so far out
#to impact the analysis.

plt.scatter(df['attend'], df['temp'])
plt.xlabel('Number of Attendance')
plt.ylabel('Temp on Game Day')
plt.title('Attendance By Temp on Game Day')
plt.grid()
plt.show()
```



In [457]:
```
#from a day of week perspective, we can see there isnt much change in the total number of attendance. Mondays
#and Wends.seems to be the
#day where there is the overall lowest number of attendance, which means there are oppurtunites to improve att
#Mondays and Wends.

plt.stem(df['day of week'], df['attend'])
plt.xlabel('Day of Week')
plt.ylabel('Total Number of Attendance')
plt.title('Attendance By Day of Week')
plt.show()
```



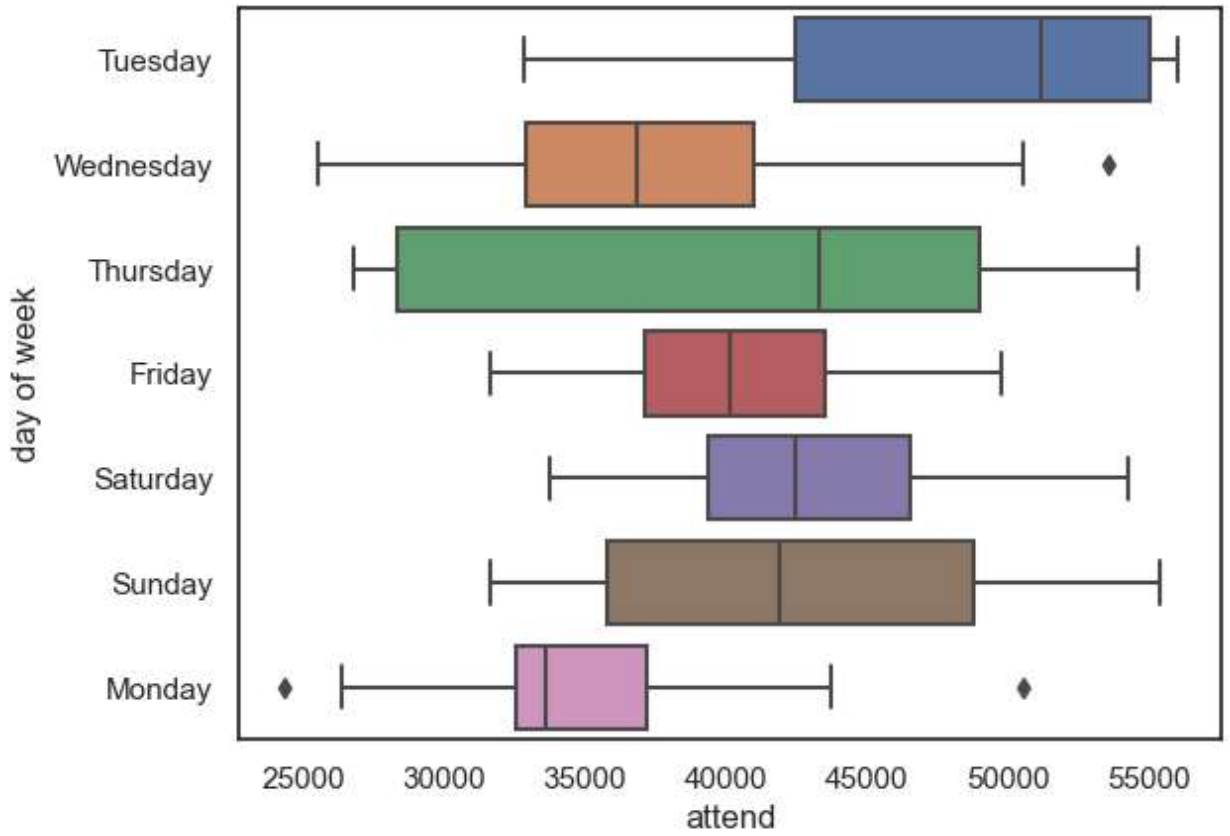
In [458]:
```
import seaborn as sns
import scipy.stats
```

In [459]:
```
#the boxplot show there are several outliers in the dataset for attendance on Wends and Monday. Again, we can
#an attendance issue on Mondays and Wends. Those days should be the focus of efforts to improve attendace on t
#during the week.

sns.boxplot(df['attend'], df['day of week'])
```

```
C:\Users\paul_\Anaconda\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following varia
bles as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passin
g other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[459]: <AxesSubplot:xlabel='attend', ylabel='day of week'>



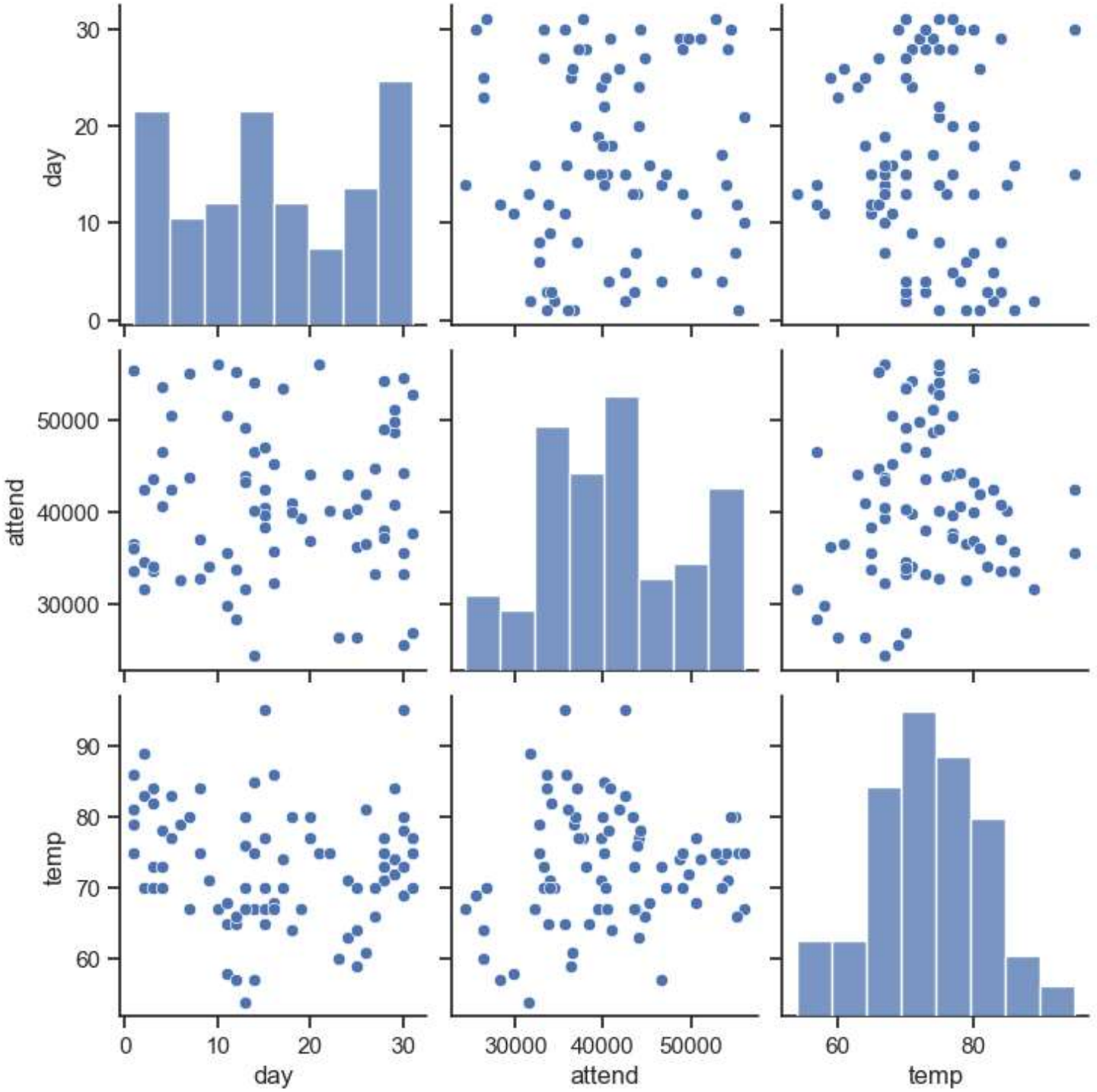
In [460]:
```
df.dtypes
```

Out[460]:
```
month           object
day              int64
attend           int64
day of week     object
opponent        object
temp             int64
skies           object
day or night    object
cap             object
shirt           object
fireworks       object
bobblehead      object
dtype: object
```

In [461]:

```
#correlative visuals that account for day of the month the game was played, attendance, and temp on game day.

sns.set(style="ticks", color_codes=True)
df= pd.read_csv('dodgers.csv')
g = sns.pairplot(df)
plt.show()
```


day
attend
temp
0
10
20
30
30000
40000
50000
60
70
80
90

In [462]:
```
#identifying correlations between day, attend, and temp.

print(df.corr())
```

```
             day    attend      temp
day     1.000000  0.027093 -0.127612
attend  0.027093  1.000000  0.098951
temp   -0.127612  0.098951  1.000000
```

In [463]:
```
#the same information as the above cell but I wanted to ensure the method was pearson. We can see there is a p
#correlation between temp and attendance where the temp drives the amount of people attending the game. The mo
#in the range of 60-90 degrees, we know more people are likely to attend the game.

df.corr(method ='pearson')
```

Out[463]:

	day	attend	temp
day	1.000000	0.027093	-0.127612
attend	0.027093	1.000000	0.098951
temp	-0.127612	0.098951	1.000000

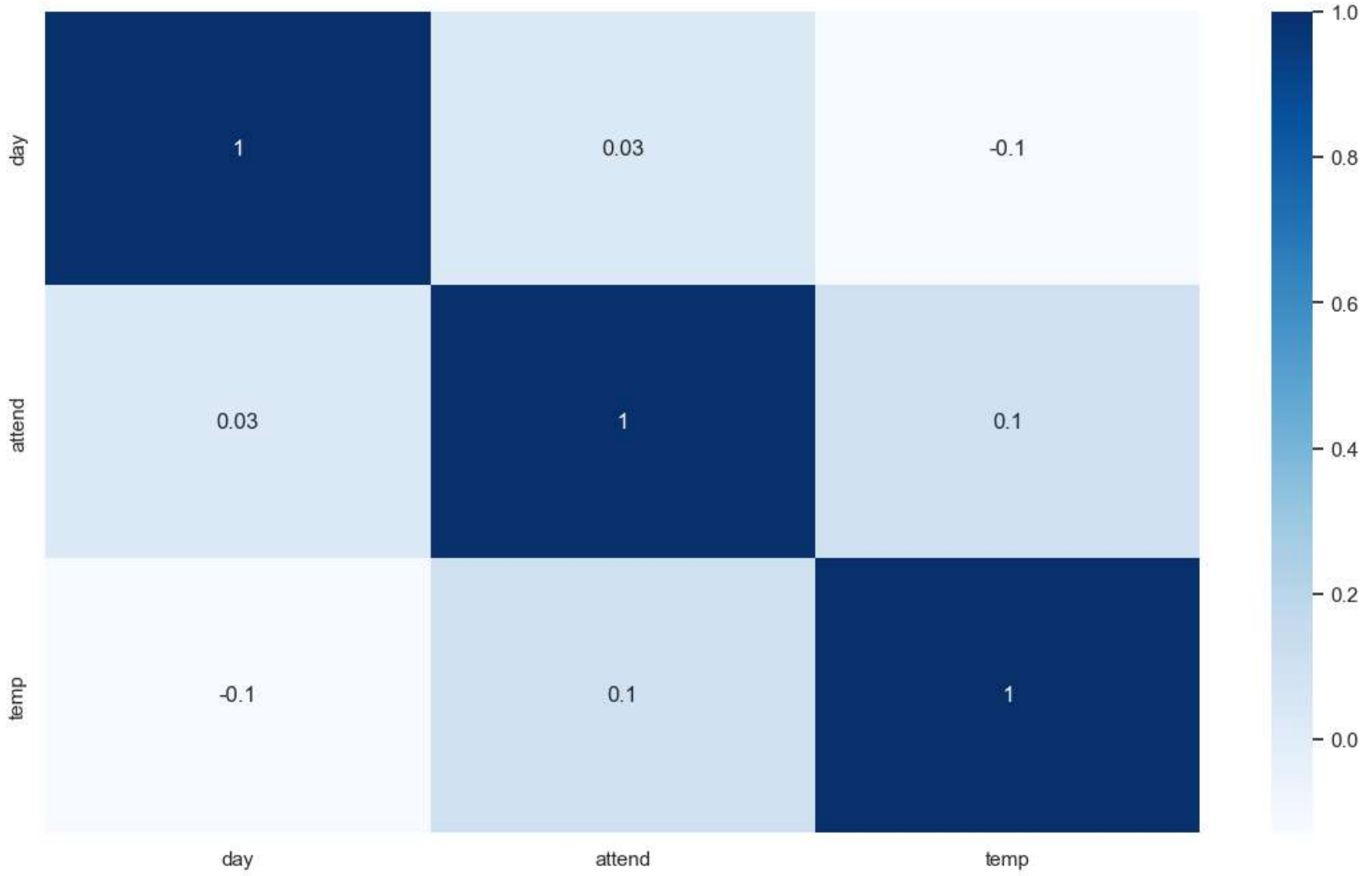
In [464]:
```
df.corr(method ='kendall')
```

Out[464]:

	day	attend	temp
day	1.000000	0.037255	-0.080142
attend	0.037255	1.000000	0.046916
temp	-0.080142	0.046916	1.000000

In [465]:
```
#the heatmap below shows the correlations between temp, attend, and day.

plt.figure(figsize=(14,8))
sns.set_theme(style="white")
corr = df.corr()
heatmap = sns.heatmap(corr, annot=True, cmap="Blues", fmt='.1g')
```



In [ ]:
```
#Final conclusions:

#There are oppurtunites to improve attendance throughout the season but in particular the team needs to look a
#Wends as the main focus of improvement efforts. We can see there are fewer people attending the games that ta
#the end of season in October and we can assume there are fewer than normal people watching the games during t
#and on Mondays and Wends. ALso, I would reccomend the team invest in more events and giveaways on Mondays and
#during the months of Sept and October in particlaur. These targeted promotions would help improve attendance
#and we know these promotions would be effective on Mondays and Wends.
#Enhancing Team Performance would be a sound investment. Investing in player development, recruitment,
#and training to improve the team's on-field performance, which may attract more fans.It's also a good idea ta
#level of fan engagment with the team and the facilities. Improving the overall amenities at the stadium would
#way to get more people to come to the games.
```

In [ ]:
```

```