In [1]:
```
#Import necessary libraries
import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
import opendatasets as od
from sklearn.model_selection import train_test_split
```

In [2]:
```
#Download the dodgers dataset from bellevue.edu
od.download("https://content.bellevue.edu/cst/dsc/630/dodgers-2022.csv")
```

```
Using downloaded and verified file: .\dodgers-2022.csv
```

In [3]:
```
#Read csv into python dataframe
dodgers_df = pd.read_csv("dodgers-2022.csv")
dodgers_df.head(5)
```

Out[3]:

	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

In [4]:
```
dodgers_df.describe()
```

Out[4]:

	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

Above we can see that the mean temperature at a game was 73.14 degrees with a mean number of attendies of 41,040,

In [5]:
```
# Summary Stats for Categorical Values
dodgers_df.describe(include=['object'])
```

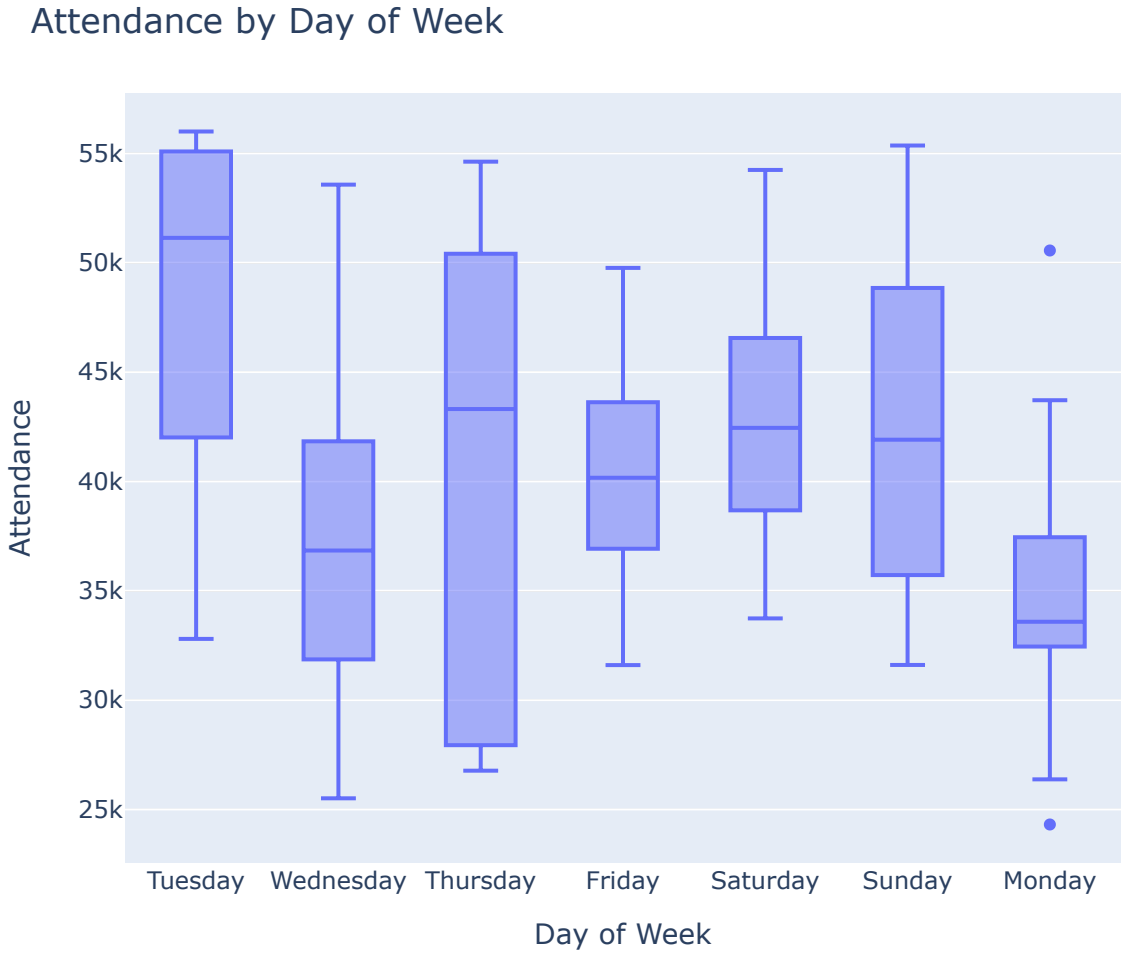
Out[5]:

	month	day_of_week	opponent	skies	day_night	cap	shirt	fireworks	bobblehead
count	81	81	81	81	81	81	81	81	81
unique	7	7	17	2	2	2	2	2	2
top	MAY	Tuesday	Giants	Clear	Night	NO	NO	NO	NO

freq	18	13	9	62	66	79	78	67	70

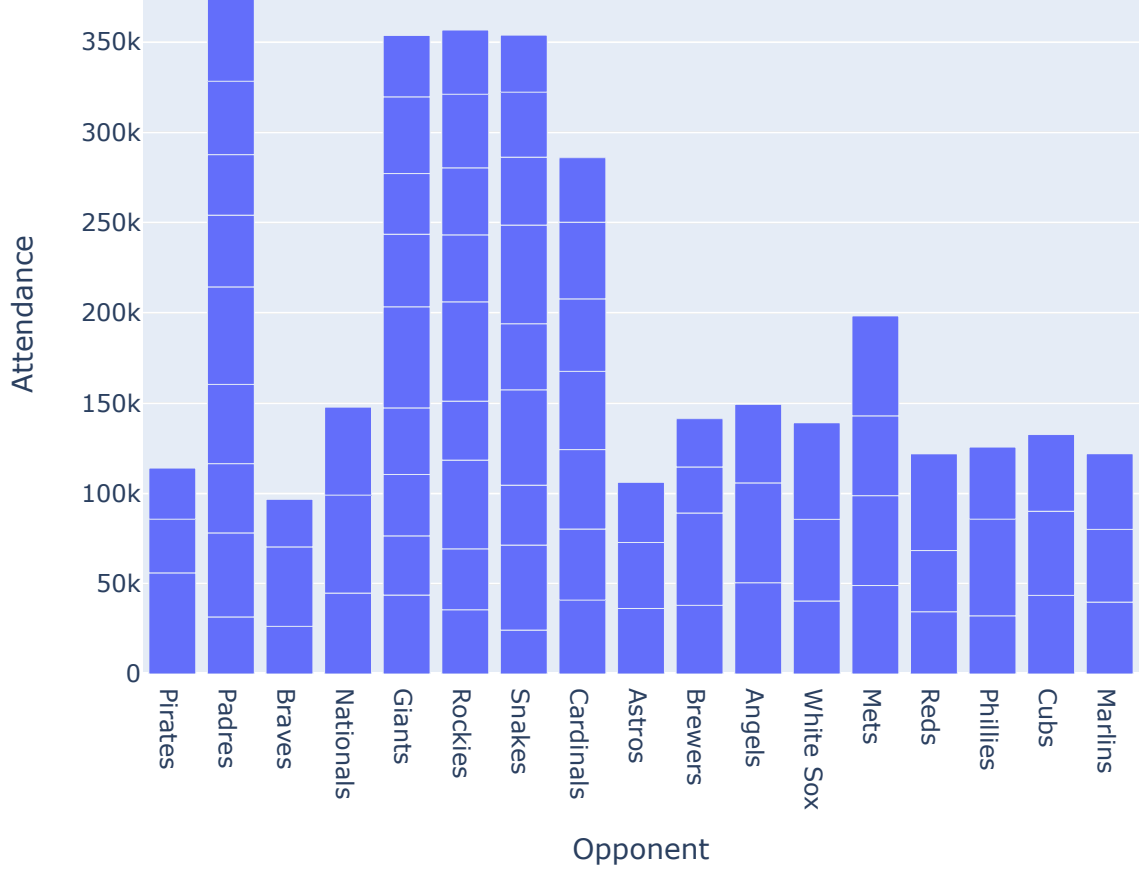
VISUALIZATIONS

In [6]:
```
#How many attended by Day of Week
fig = px.box(dodgers_df, y="attend", x="day_of_week")
fig.update_layout(
    xaxis_title="Day of Week",
    yaxis_title="Attendance",
    title="Attendance by Day of Week")
fig.show('notebook')
```



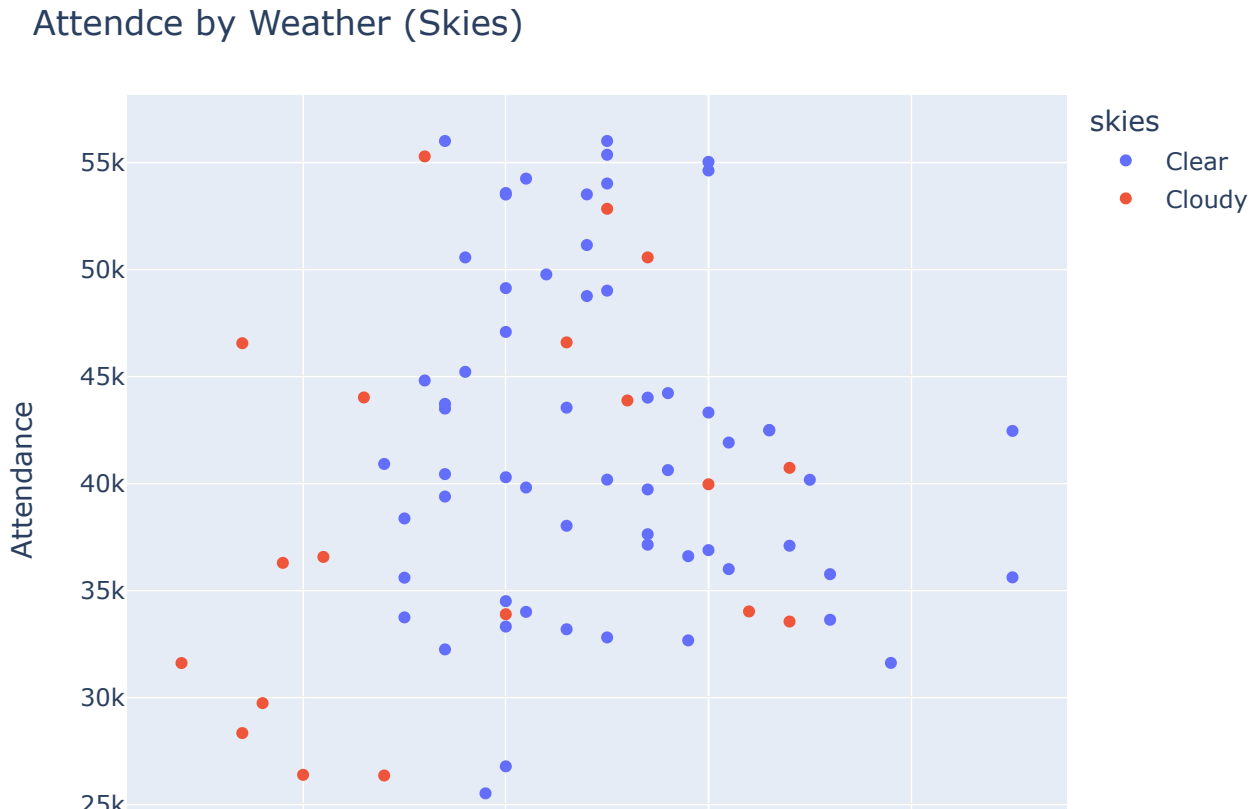
From the chart above, we can see that maximum attendance is on Tuesdays and least attendance on Monday.

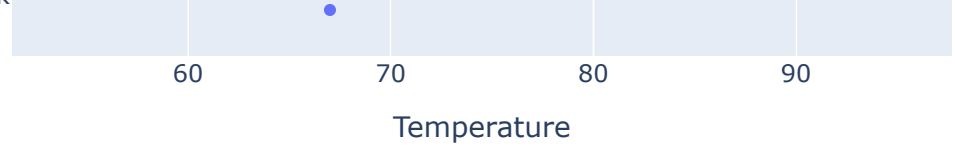
In [7]:
```
#Total number of attendes by Opponent
fig = px.bar(dodgers_df, y="attend", x="opponent")
fig.update_layout(
    xaxis_title="Opponent",
    yaxis_title="Attendance",
    title="Attendance by Opponent")
fig.show('notebook')
```



Based on the above plot, playing with the opponents Padres had maximum attendance.

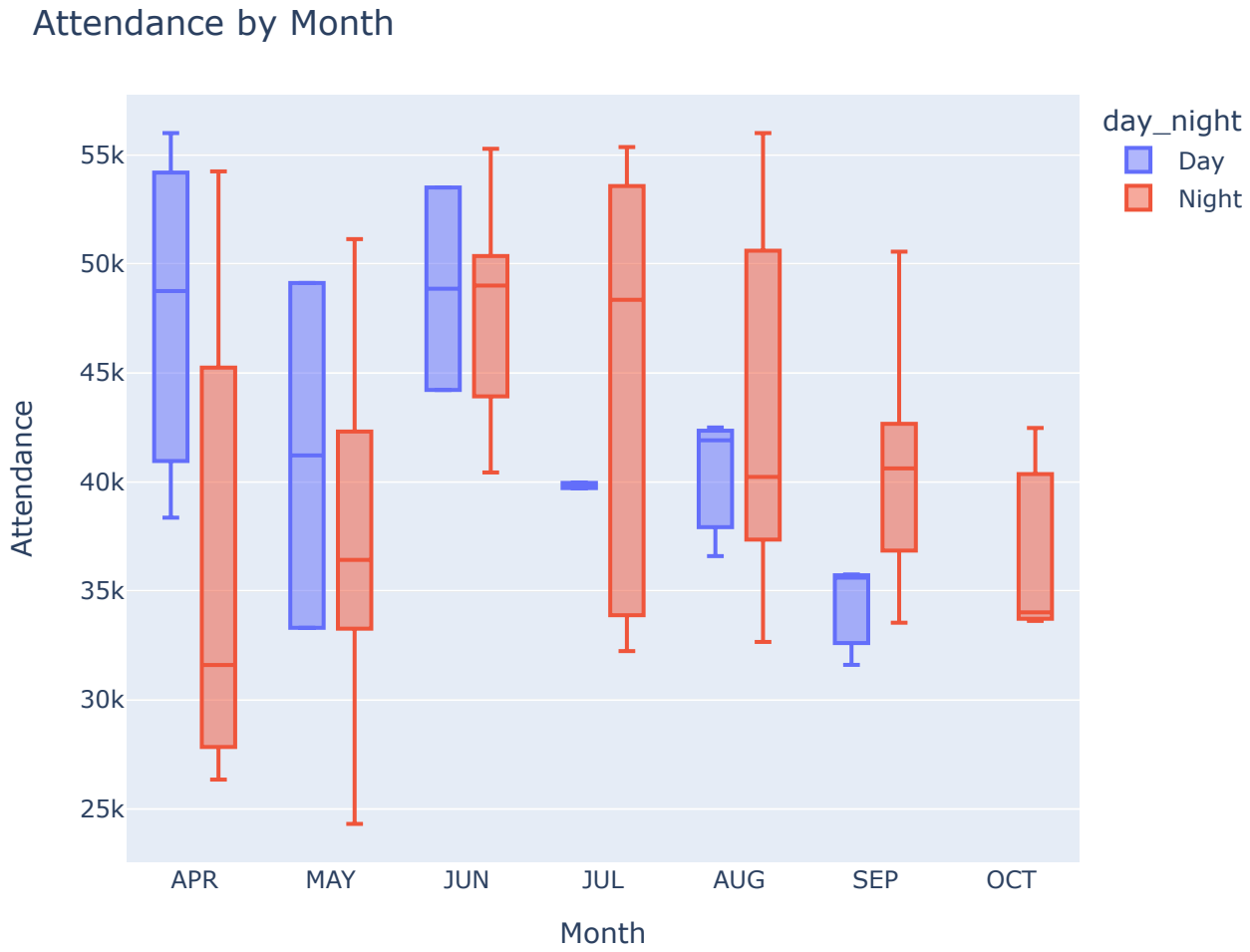
In [14]:
```
#Attendes vs Weather
fig = px.scatter(dodgers_df, y="attend", x="temp",color='skies')
fig.update_layout(
    xaxis_title="Temperature",
    yaxis_title="Attendance",
    title="Attendce by Weather (Skies)")
fig.show('notebook')
```





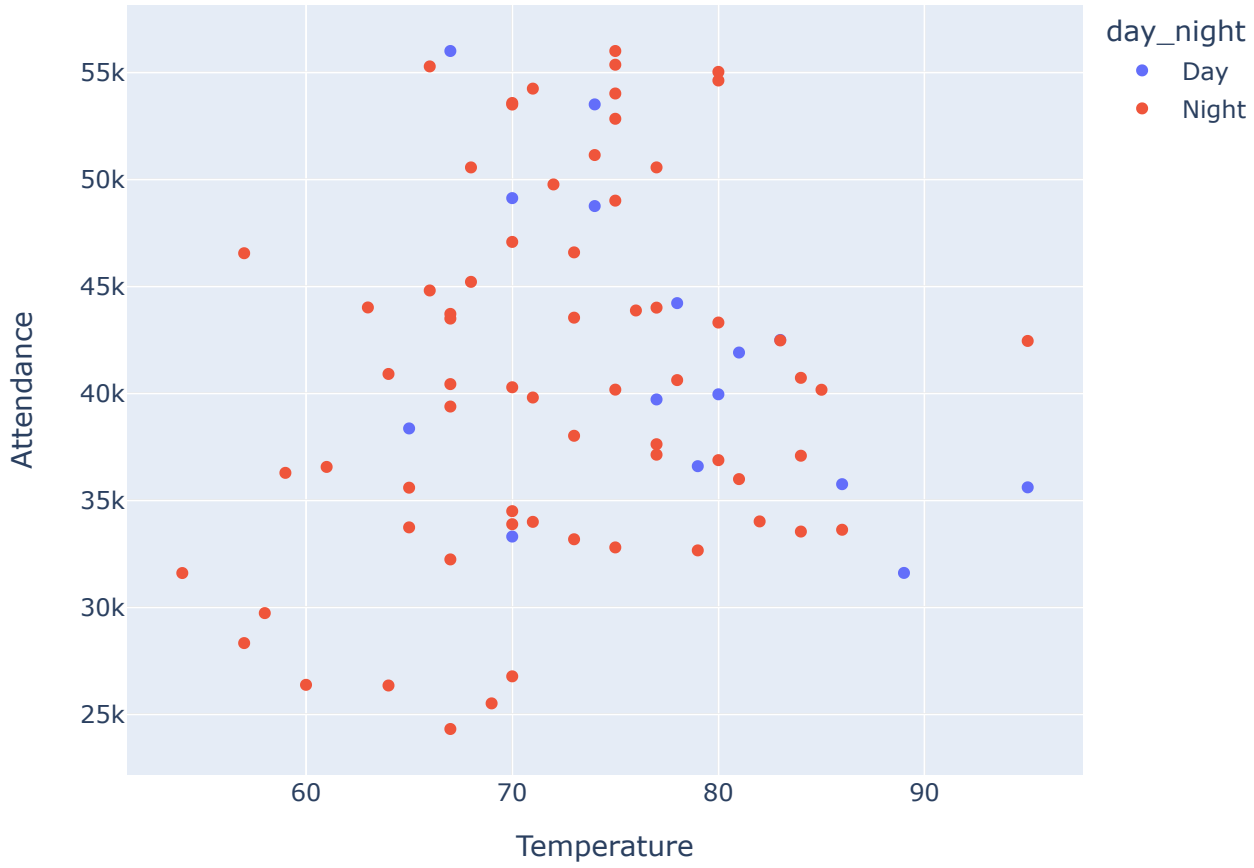
From the above chart, we can see that the ideal temperature for maximum attendance is 75 with clear skies.

In [15]:
```
#How many attended by Month
fig = px.box(dodgers_df, y="attend", x="month",color = "day_night")
fig.update_layout(
    xaxis_title="Month",
    yaxis_title="Attendance",
    title="Attendance by Month")
fig.show('notebook')
```



April has the maximum attendes, preferably in the day.

In [ ]:
```
#Attendes vs Weather
fig = px.scatter(dodgers_df, y="attend", x="temp",color='day_night')
fig.update_layout(
    xaxis_title="Temperature",
    yaxis_title="Attendance",
    title="Attendce by Weather (Day/Night)")
fig.show('notebook')
```



The maximum attendance is in the night when the temparature is in 70's.

CORRELATION

In [ ]:
```
# To find the correlation among
# the columns using pearson method
dodgers_df.corr(method ='pearson',numeric_only=True)
```

Out[ ]:

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

The above correlation matrix shows the relationship between the numerical values. Day has a negative correlation with temperature but this is not significant (0.1). There is no strong correlation (positive/negative) between any numeric features.

Let's see the relation between the categorical and non-categorical (numeric) variables using Spearman correlation matrix

In [ ]:
```
# To support the Spearman Correlation Matrix, create dummy variables for the object type
df = pd.concat([dodgers_df.drop(['month', 'day_of_week', 'opponent','skies','day_night',
df.head(5)
```

Out[ ]:

	day	attend	temp	month_APR	month_AUG	month_JUL	month_JUN	month_MAY	month_OCT	month_SEP

0	10	56000	67	True	False	False	False	False	False	False
1	11	29729	58	True	False	False	False	False	False	False
2	12	28328	57	True	False	False	False	False	False	False
3	13	31601	54	True	False	False	False	False	False	False
4	14	46549	57	True	False	False	False	False	False	False

5 rows × 44 columns

In [ ]:
```
df.corr('spearman').style.background_gradient(cmap="Blues")
```

Out[ ]:

	day	attend	temp	month_APR	month_AUG	month_JUL	month_JUN	mon
day	1.000000	0.063626	-0.123692	0.104875	-0.028569	-0.079586	0.108461	0
attend	0.063626	1.000000	0.090628	-0.055739	0.101270	0.096614	0.314192	-0
temp	-0.123692	0.090628	1.000000	-0.495820	0.296848	0.012656	-0.132964	-0
month_APR	0.104875	-0.055739	-0.495820	1.000000	-0.198811	-0.173913	-0.147442	-0
month_AUG	-0.028569	0.101270	0.296848	-0.198811	1.000000	-0.198811	-0.168550	-0
month_JUL	-0.079586	0.096614	0.012656	-0.173913	-0.198811	1.000000	-0.147442	-0
month_JUN	0.108461	0.314192	-0.132964	-0.147442	-0.168550	-0.147442	1.000000	-0
month_MAY	0.153172	-0.223536	-0.337159	-0.222911	-0.254824	-0.222911	-0.188982	1
month_OCT	-0.293820	-0.109043	0.268880	-0.081786	-0.093495	-0.081786	-0.069338	-0
month_SEP	-0.113057	-0.109991	0.527833	-0.173913	-0.198811	-0.173913	-0.147442	-0
day_of_week_Friday	0.134612	-0.030209	-0.167878	0.007013	0.051309	-0.087664	0.059456	0
day_of_week_Monday	-0.119007	-0.325514	-0.024568	-0.076087	-0.019881	0.119565	-0.036860	0
day_of_week_Saturday	0.083503	0.128028	-0.044672	0.007013	-0.035275	-0.087664	0.059456	0
day_of_week_Sunday	0.035273	0.051787	0.237768	0.007013	-0.035275	0.007013	-0.047565	0
day_of_week_Thursday	0.172376	-0.008776	0.014286	0.037438	0.009782	-0.106966	0.072548	-0
day_of_week_Tuesday	-0.090701	0.333736	-0.020895	0.007013	-0.035275	0.101690	-0.047565	0
day_of_week_Wednesday	-0.165867	-0.167959	0.010423	0.021739	0.069584	0.021739	-0.036860	-0
opponent_Angels	-0.106335	0.204106	-0.184855	-0.081786	-0.093495	-0.081786	0.554700	-0
opponent_Astros	0.179090	-0.156575	-0.226868	-0.081786	-0.093495	-0.081786	-0.069338	0
opponent_Braves	0.141313	-0.167758	-0.278683	0.470270	-0.093495	-0.081786	-0.069338	-0
opponent_Brewers	0.319518	-0.134038	-0.059812	-0.095050	-0.108657	-0.095050	-0.080582	0
opponent_Cardinals	0.038556	0.015034	0.181659	-0.128262	-0.146625	-0.128262	-0.108740	0
opponent_Cubs	-0.237854	0.109043	0.082625	-0.081786	0.411377	-0.081786	-0.069338	-0
opponent_Giants	-0.216080	-0.086529	0.196922	-0.147442	0.134840	-0.147442	-0.125000	0
opponent_Marlins	0.159502	0.002796	0.032210	-0.081786	0.411377	-0.081786	-0.069338	-0
opponent_Mets	0.130490	0.248580	0.076901	-0.095050	-0.108657	0.065347	0.463348	-0
opponent_Nationals	0.225262	0.204106	-0.079824	0.470270	-0.093495	-0.081786	-0.069338	-0
opponent_Padres	-0.188335	0.038644	-0.010099	0.184302	-0.168550	0.184302	-0.125000	-0

opponent_Phillies	0.053167	-0.011184	-0.025208	-0.081786	-0.093495	0.470270	-0.069338	-0
opponent_Pirates	-0.131519	-0.082481	-0.273081	0.470270	-0.093495	-0.081786	-0.069338	-0
opponent_Reds	-0.264438	-0.030756	-0.092428	-0.081786	-0.093495	0.470270	-0.069338	-0
opponent_Rockies	-0.021860	-0.082328	0.161577	-0.147442	0.134840	-0.147442	-0.125000	0
opponent_Snakes	0.052969	-0.089049	0.167468	-0.147442	0.134840	0.073721	-0.125000	0
opponent_White Sox	0.029382	0.139799	-0.102230	-0.081786	-0.093495	-0.081786	0.554700	-0
skies_Clear	0.054252	0.144553	0.259024	-0.343251	0.188903	-0.097204	0.103011	0
skies_Cloudy	-0.054252	-0.144553	-0.259024	0.343251	-0.188903	0.097204	-0.103011	-0
day_night_Day	0.052377	0.031944	0.249189	0.069584	0.018182	-0.019881	0.033710	-0
day_night_Night	-0.052377	-0.031944	-0.249189	-0.069584	-0.018182	0.019881	-0.033710	0
cap_NO	0.194109	0.051039	-0.066466	0.066354	-0.128951	-0.157591	0.056254	0
cap_YES	-0.194109	-0.051039	0.066466	-0.066354	0.128951	0.157591	-0.056254	-0
shirt_NO	0.037777	-0.139799	-0.011203	-0.102233	0.093495	0.081786	-0.138675	0
shirt_YES	-0.037777	0.139799	0.011203	0.102233	-0.093495	-0.081786	0.138675	-0
fireworks_NO	-0.091546	-0.015361	0.178363	0.006808	-0.034245	0.006808	-0.046176	0
fireworks_YES	0.091546	0.015361	-0.178363	-0.006808	0.034245	-0.006808	0.046176	-0

*From Spearman's Correlation Matrix above

- When 2 variables move in the same direction, i.e. when one variable increases the other variable also increases or vice-versa, the relationsip between the two variables is said to be a positive correlation.
- When 2 variables move in the same direction, i.e. when one variable increases the other variable also decreases or vice-versa, the relationsip between the two variables is said to be a negative correlation.

OBSERVATIONS:

Based on the above Spearman's Correlation Matrix, following are the features that have a positive correlation.

MONTHS: April, May and June months have more attendance than other months, which implies people prefer attending games in the summer months.

DAY OF WEEK: Based on the correlation, Tuedays and Saturdays have a positive correlation in comparison to other days.

OPPONENTS: The most attendance is seen when Dodgers played Mets, Angels, Nationals and WhiteSox.

DAY/NIGHT: Skies: Better attendance is seen in the Day with Clear skies

SHIRTS: Giving free shirts seems to have a positive correlation with the number of attendes.

Recommendations to Management:

Based on the above observations, management can conduct more matches in summer months, preferably on Tuesdays and Saturdays in the Day with assumed clear skies. To drive more attendance, the management can arrange for free shirts to the fans.

Assumption:

Weather conditions in these months may not always be clear.

References:

https://en.wikipedia.org/wiki/Los_Angeles_Dodgers

Since I am not a sports fan, I had to read a little about Dodgers.