Assignment 3.2: Using Data to Improve MLB Attendance

For the assignment, the goal is to analyze the data and provide a recommendation on how to improve attendance.

In [41]:
```
# importing libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [42]:
```
# importing warning to surpress the future warnings
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

In [43]:
```
# reading dataset
df = pd.read_csv('Datasets/dodgers-2022.csv')
```

In [44]:
```
# reviewing first 10 rows
df.head(10)
```

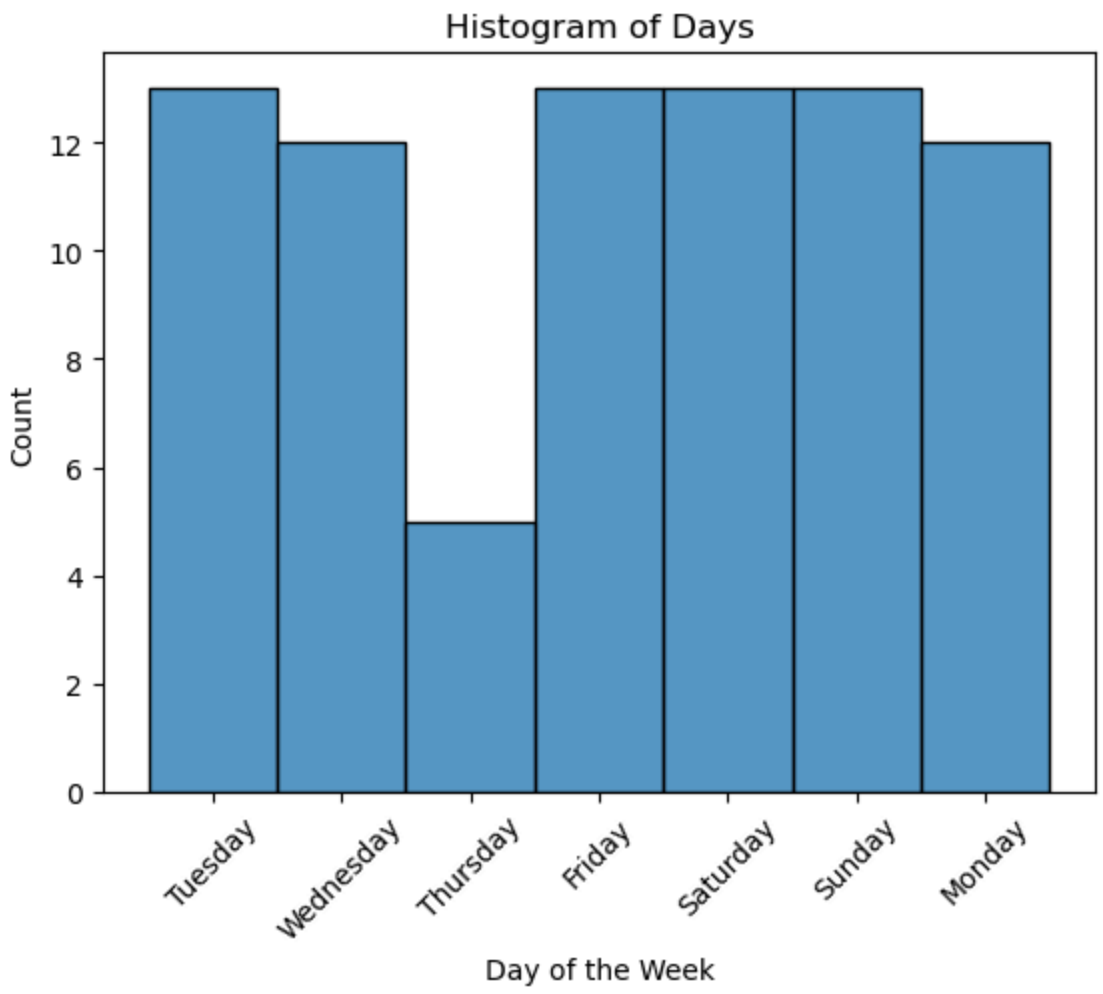
Out[44]:

	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

In [75]:
```
# reviewing shape of the dataframe
df.shape
```

Out[75]: (81, 12)

In [70]:
```
# creating a histogram of days of the weeks
sns.histplot(data=df, x='day_of_week')
plt.xlabel('Day of the Week')
plt.xticks(rotation = 45)
plt.title('Histogram of Days')
plt.show()
```

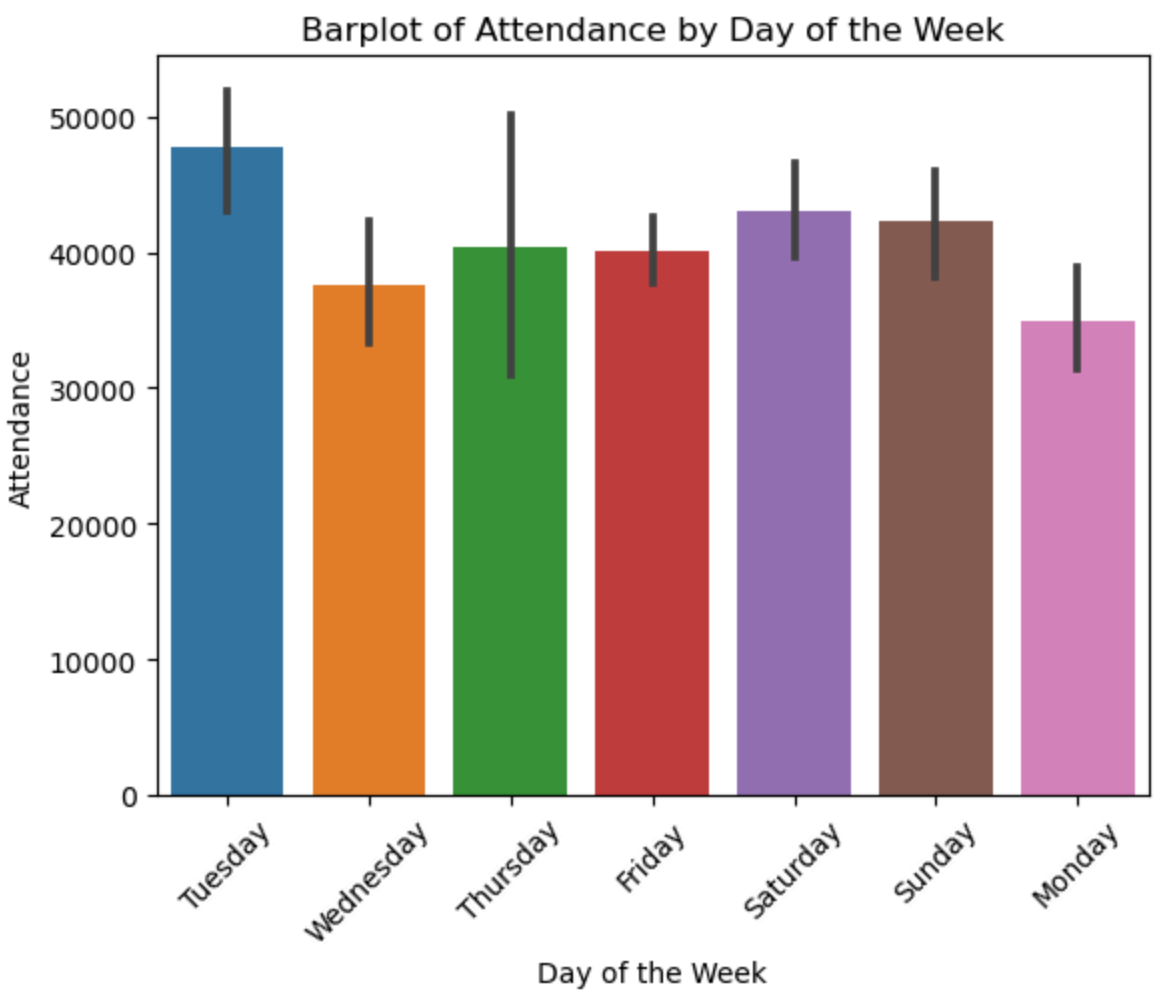


Reviewing the days of the week with a histogram can tell us which days appear the most in the dataframe.

Barplots of Attendance

In [71]:
```
# creating a barplot of attendance vs days of the week
sns.barplot(x = 'day_of_week', y = 'attend', data = df)
```

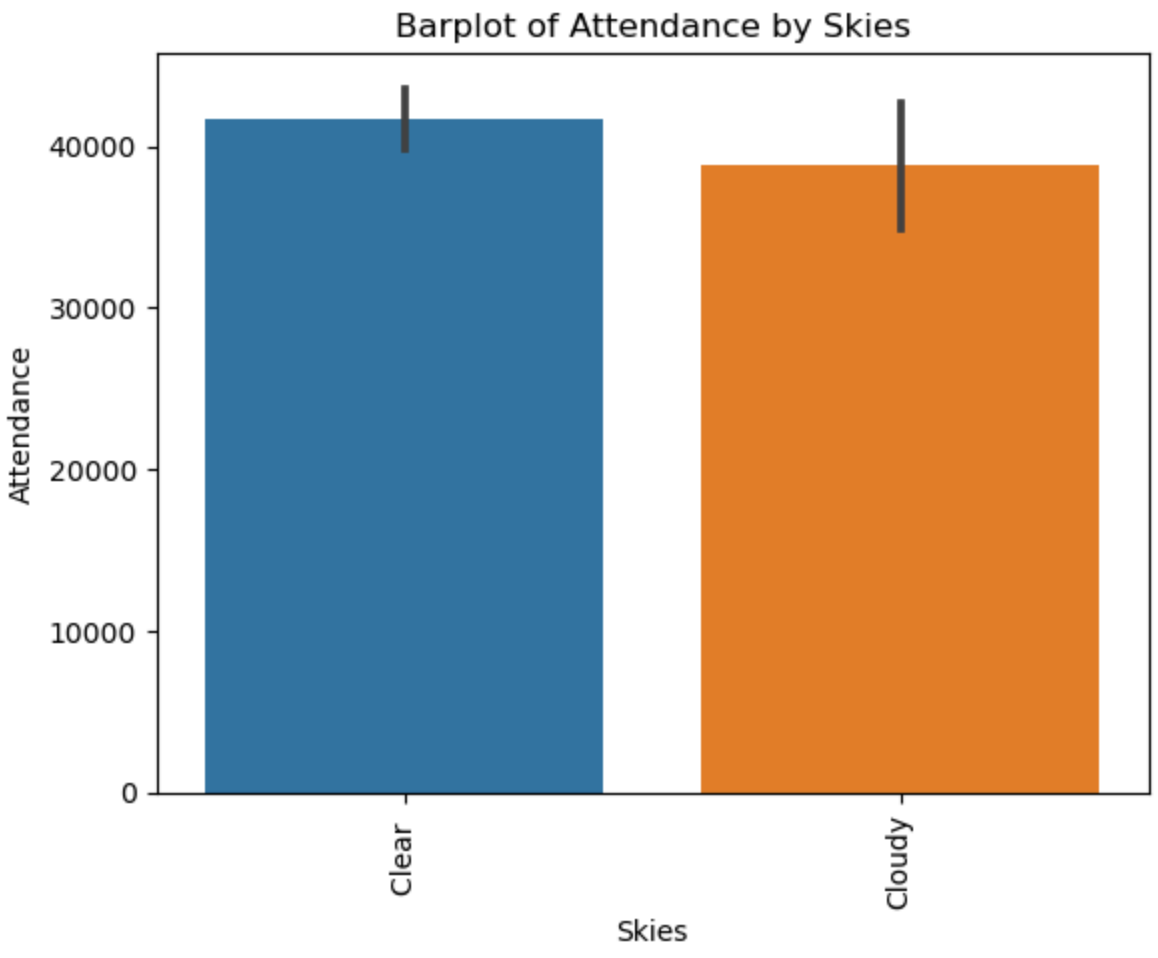
```
plt.xlabel('Day of the Week')
plt.ylabel('Attendance')
plt.xticks(rotation = 45)
plt.title('Barplot of Attendance by Day of the Week')
plt.show()
```



Reviewing the attendance by day of the week barplot shows a surprising find, I would expect that the weekend days would have higher attendance than days during the traditional work week, and they mostly do. However, Tuesday games are showing a higher attendance than any other day.

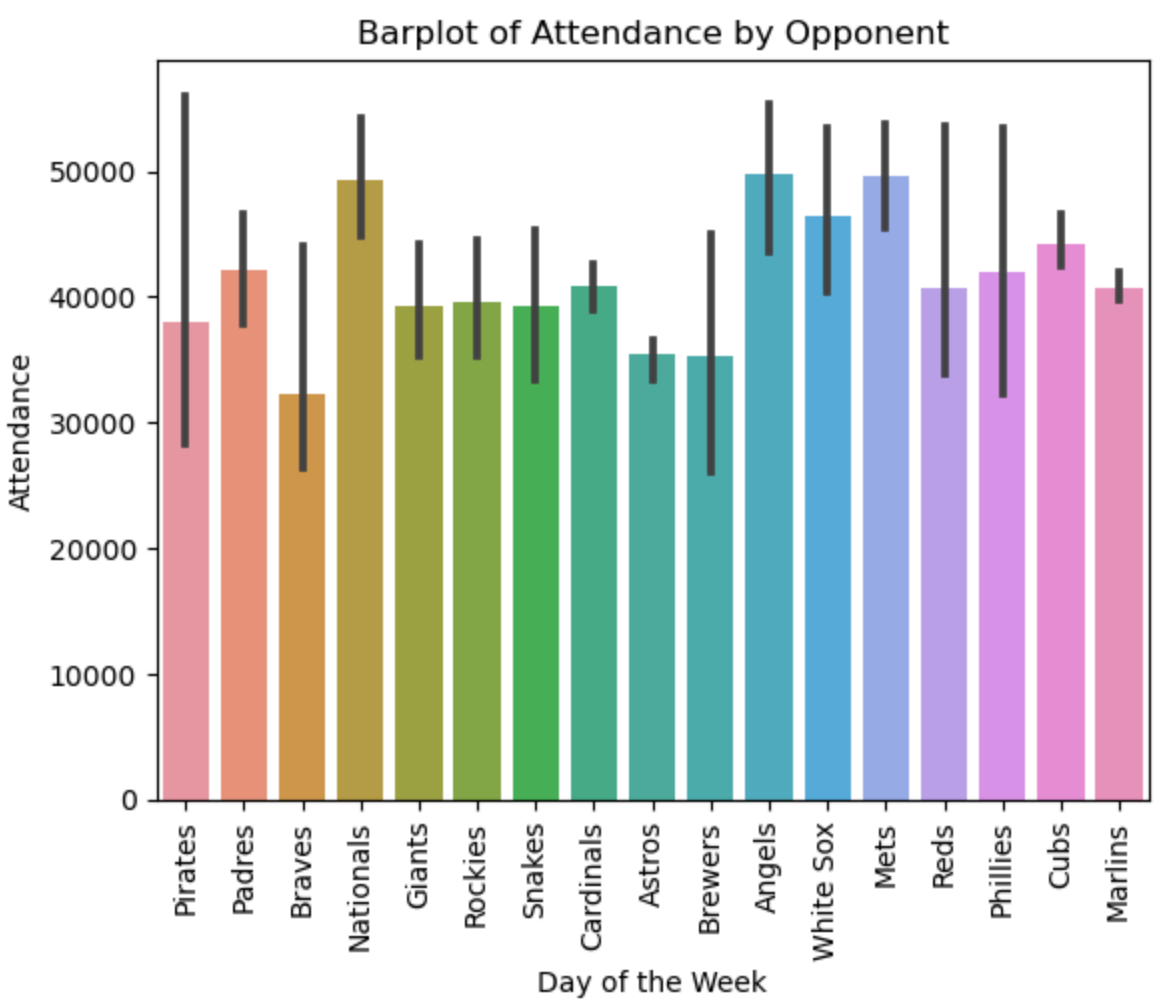
This creates an interest in finding why Tueday is the most attended day.

In [72]:
```
sns.barplot(x = 'skies', y = 'attend', data = df)
plt.xlabel('Skies')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Skies')
plt.show()
```



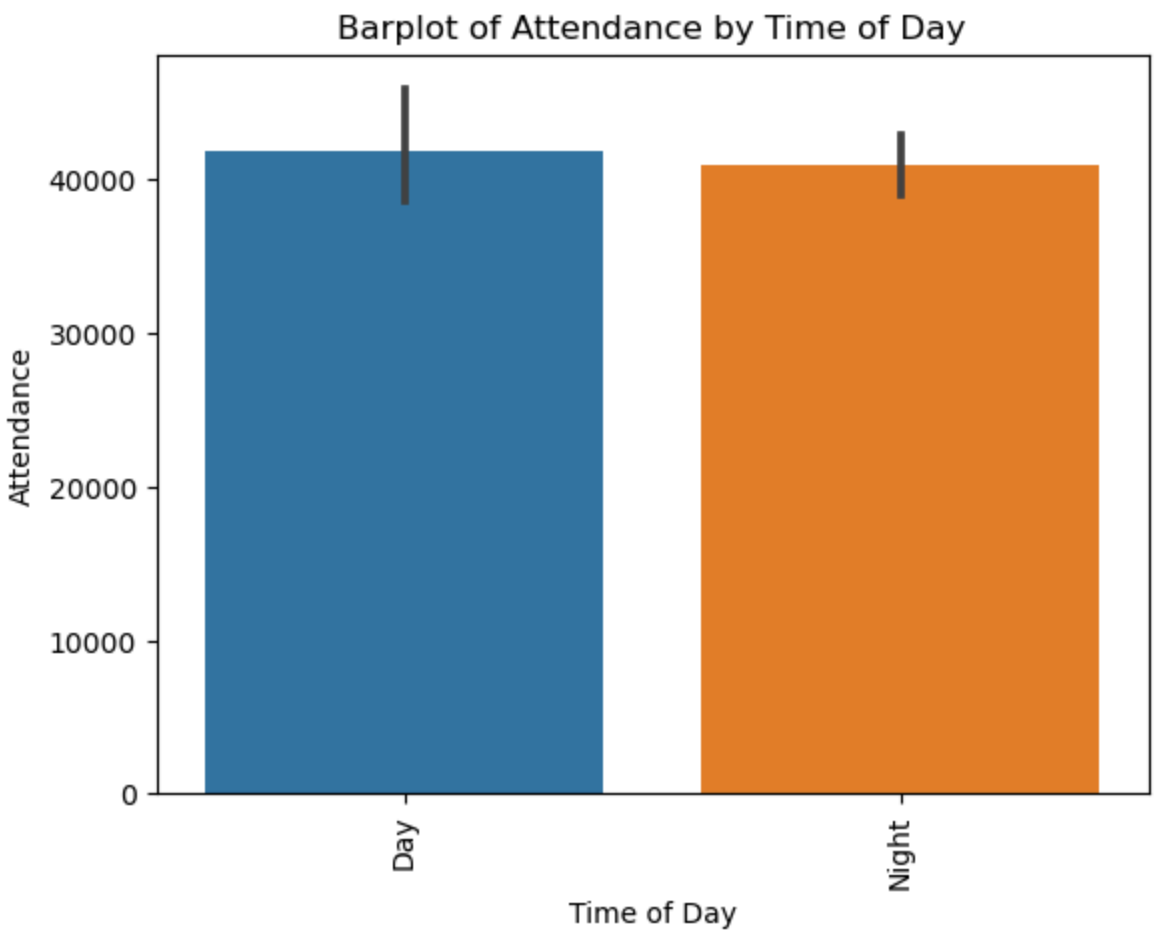
Attendance by the type of sky, shows an expected result. Clear skies have higher attendance than cloudy skies.

In [73]:
```
sns.barplot(x = 'opponent', y = 'attend', data = df)
plt.xlabel('Day of the Week')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Opponent')
plt.show()
```



The attendance by opponent barplot does not appear to provide any insights itself, it does offer more questions and room for analysis though. It would be intersting to analyze what the highest attendace opponents have in common. Perhaps they are geographiclly close to the home team?

In [74]:
```
# creating a barplot of attendance by time, day or night
sns.barplot(x = 'day_night', y = 'attend', data = df)
plt.xlabel('Time of Day')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Time of Day')
plt.show()
```



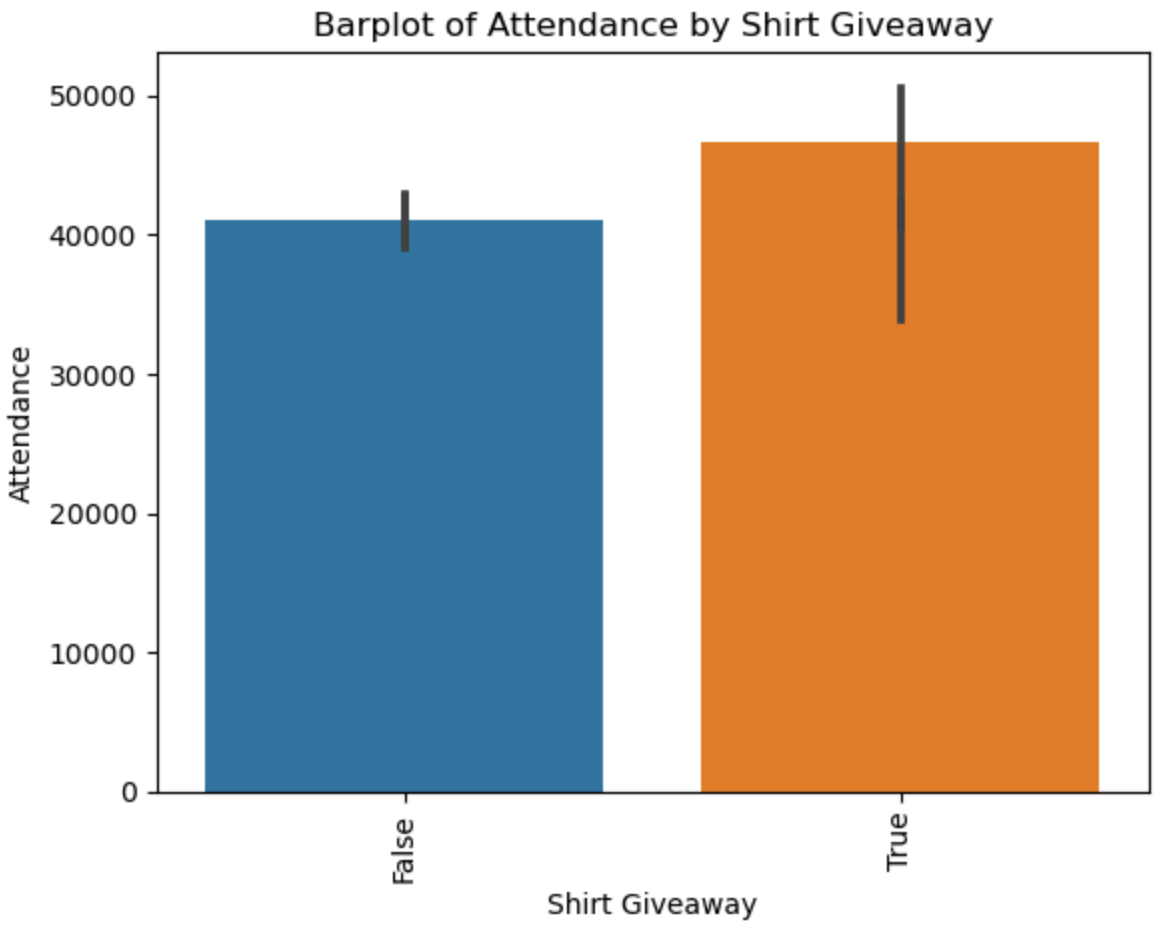
The barplot of attendance by time of day does not appear to offer any significant insights. The attendance difference between day and night games appears to be minimal.

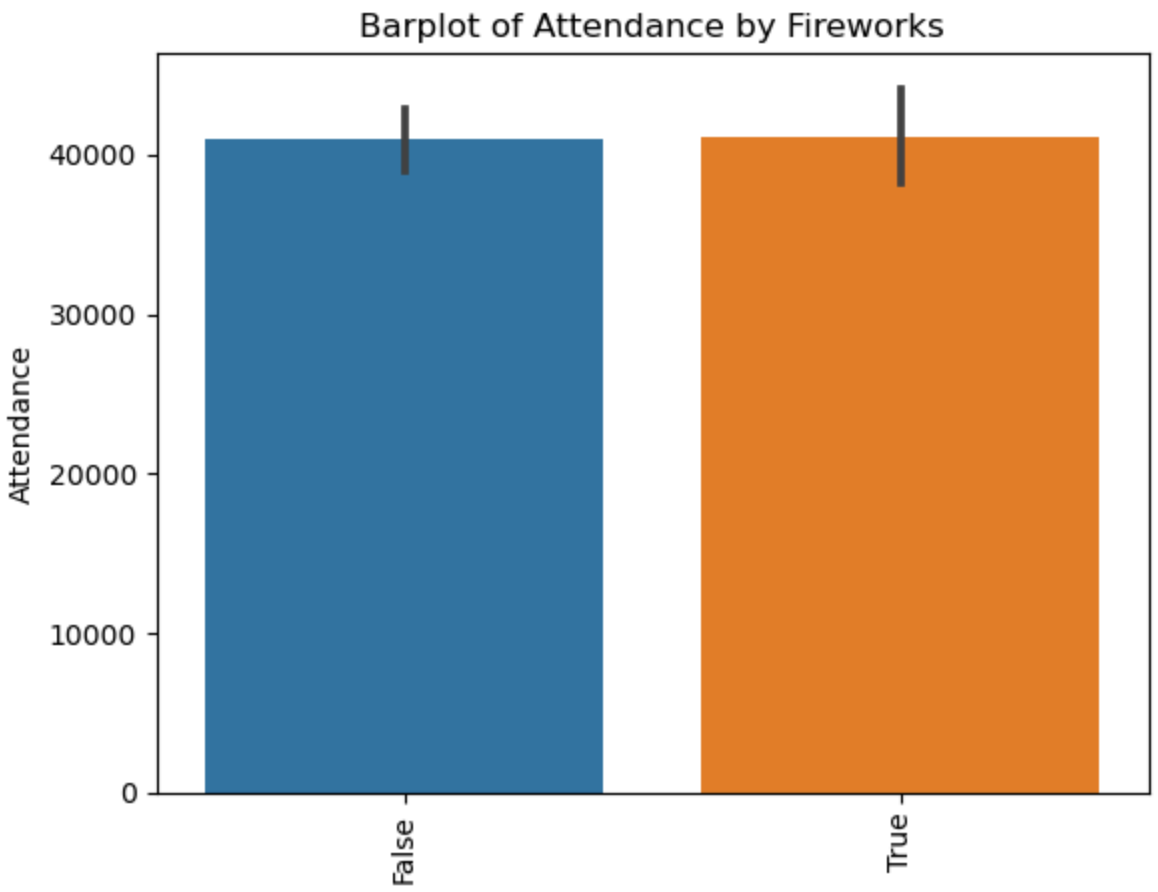
Giveaways

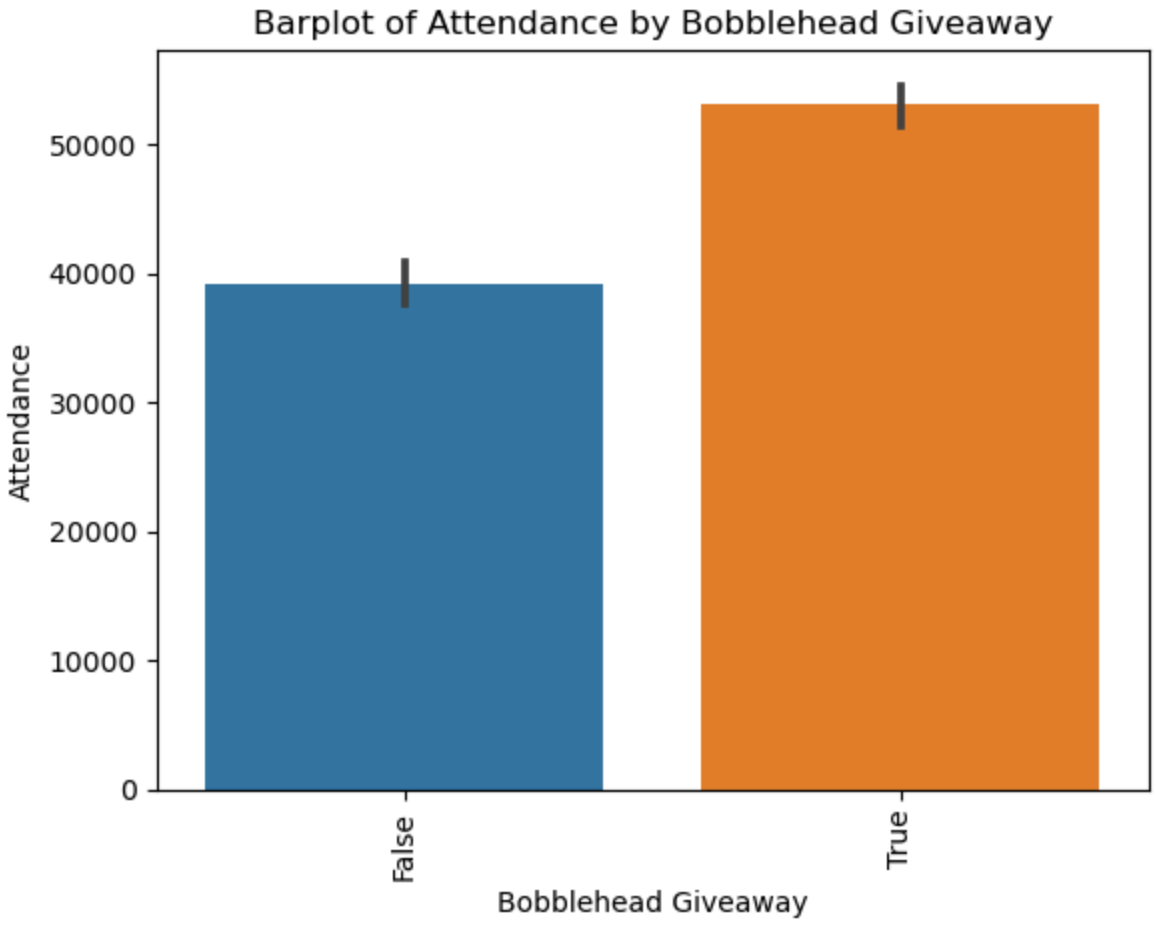
I decided to group all of the plots for giveaways together to evaulate them all together.

In [89]:
```
# creating bar plots for the various giveaways by attendance
sns.barplot(x = 'cap', y = 'attend', data = df)
plt.xlabel('Cap Giveaway')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Cap Giveaway')

sns.barplot(x = 'shirt', y = 'attend', data = df)
plt.xlabel('Shirt Giveaway')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Shirt Giveaway')
plt.show()

sns.barplot(x = 'fireworks', y = 'attend', data = df)
```

```
plt.xlabel('Fireworks')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Fireworks')
plt.show()

sns.barplot(x = 'bobblehead', y = 'attend', data = df)
plt.xlabel('Bobblehead Giveaway')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Bobblehead Giveaway')
plt.show()
```







The bar plots above show that the givewawy related with the highest attendance is the Bobblehead giveaway. This is once again surprising, at least to me. I would expect shirts or hats to be more popular since it seems they would be more widely popular.

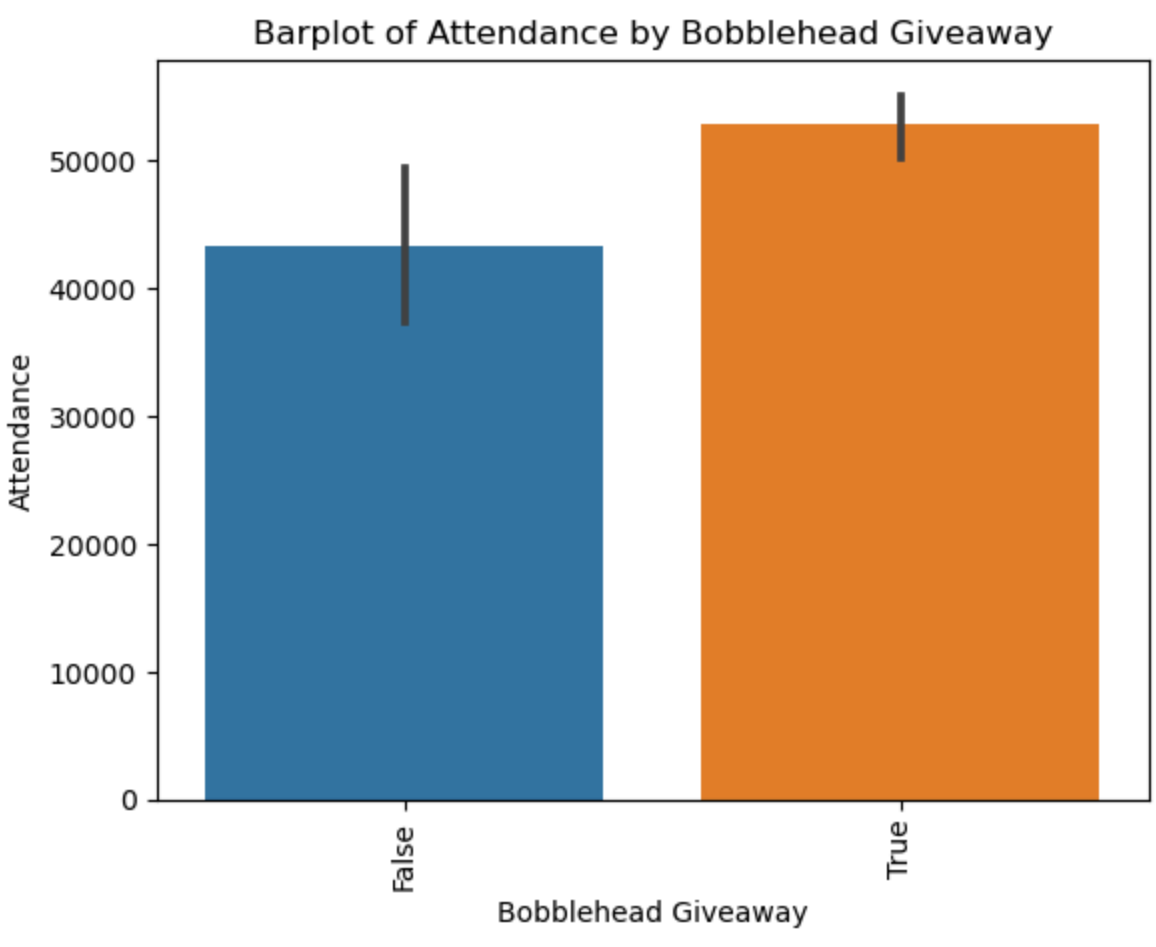
In [77]:
```
# filtering the dataframe by day of the week being Tuesday
df_tues = df[df['day_of_week'] == 'Tuesday']
df_tues
```

Out[77]:

	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks	bobblehead
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	False	False	False	False
7	APR	24	44014	Tuesday	Braves	63	Cloudy	Night	False	False	False	False
13	MAY	8	32799	Tuesday	Giants	75	Clear	Night	False	False	False	False
19	MAY	15	47077	Tuesday	Snakes	70	Clear	Night	False	False	False	True
27	MAY	29	51137	Tuesday	Brewers	74	Clear	Night	False	False	False	True
31	JUN	12	55279	Tuesday	Angels	66	Cloudy	Night	False	False	False	True
41	JUL	3	33884	Tuesday	Reds	70	Cloudy	Night	True	False	False	False
47	JUL	17	53498	Tuesday	Phillies	70	Clear	Night	False	False	False	False
50	JUL	31	52832	Tuesday	Snakes	75	Cloudy	Night	False	False	False	True
56	AUG	7	55024	Tuesday	Rockies	80	Clear	Night	False	False	False	True
59	AUG	21	56000	Tuesday	Giants	75	Clear	Night	False	False	False	True
69	SEP	4	40619	Tuesday	Padres	78	Clear	Night	False	True	False	False
79	OCT	2	42473	Tuesday	Giants	83	Clear	Night	False	False	False	False

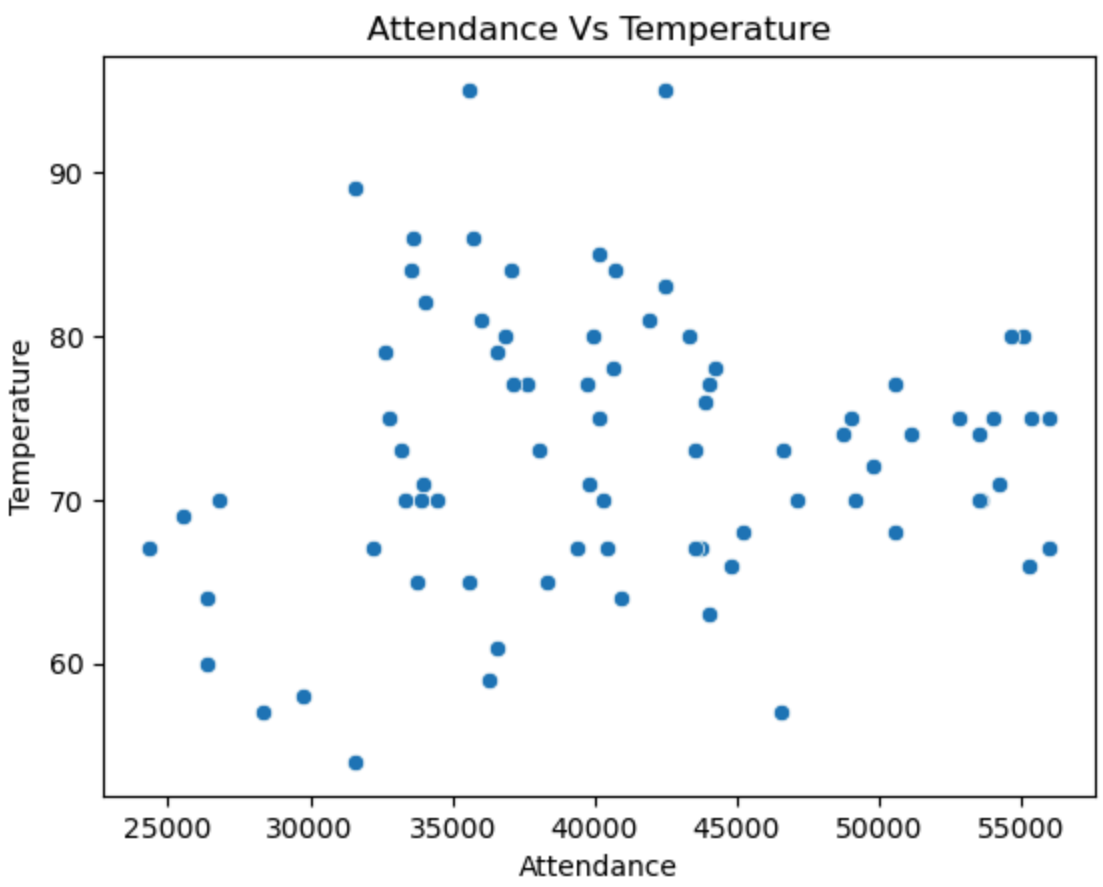
The dataframe was filtered to only show data from the day 'Tuesday' to try and find anything in the dataframe that may explain the high attendance on Tuesdays.

In [90]:
```
sns.barplot(x = 'bobblehead', y = 'attend', data = df_tues)
plt.xlabel('Bobblehead Giveaway')
plt.ylabel('Attendance')
plt.xticks(rotation = 90)
plt.title('Barplot of Attendance by Bobblehead Giveaway')
plt.show()
```



After filtering the dataframe to Tuesday only data, another barplot was created to view the relationship between the bobblehead giveaway and attendance. The highest attendance on tuesday were with bobble head giveaways.

In [50]:
```
# creating a scatterplot of temp and attendance
sns.scatterplot(x='attend', y='temp', data=df)
plt.xlabel('Attendance')
plt.ylabel('Temperature')
plt.title('Attendance Vs Temperature')
plt.show()
```



Based on the scatterplot, there does not appear to be a correlation between attendance and temperature.

Summary

In reviewing the data for MLB attendance a few insights were gained. I started with creating and reviewing a histogram of the days. This showed that there was not a single day that appeared to be significantly overrepresented.

Several bar plots were then created with the focus being attendance. After reviewing the bar plots it was found that the Tuesday games had the most attendance. This came as a surprise, with a traditional work week typically Monday through Friday, I was expecting the weekend days to show higher attendance numbers than any weekdays.

Bar plots of giveaways was the next step in trying to identify factors the influenced attendance. Another surprise was found when reviewing the bar plots. It appears that the bobblehead giveaways were related to higher attendance

Based on the results of the analysis, my suggestion would be to focus on the under performing attendance days and offer bobblehead giveaways on those days. By doing so the attendance on those under performing days should show an increase.