DSC630

Xin Tang

Week3 assignment: Using data to improve MLB Attendance

Data description and Goal

This datasets contains the info of LA Dodgers MLB game attendance data, the goal is to analysis the data and give recommendation of actions to improve attendance.

In [2]:
```
# First step is to load data,  analysis and find out data distribution and relationship
import numpy as np
import pandas as pd

# Loading the data
df = pd.read_csv("dodgers-2022.csv")

# Check data load correctly
print(df.head(2))
```

```
  month  day  attend day_of_week opponent  temp   skies day_night cap shirt  \
0   APR   10   56000     Tuesday  Pirates    67  Clear        Day  NO    NO
1   APR   11   29729   Wednesday  Pirates    58  Cloudy     Night  NO    NO

  fireworks bobblehead
0        NO         NO
1        NO         NO
```

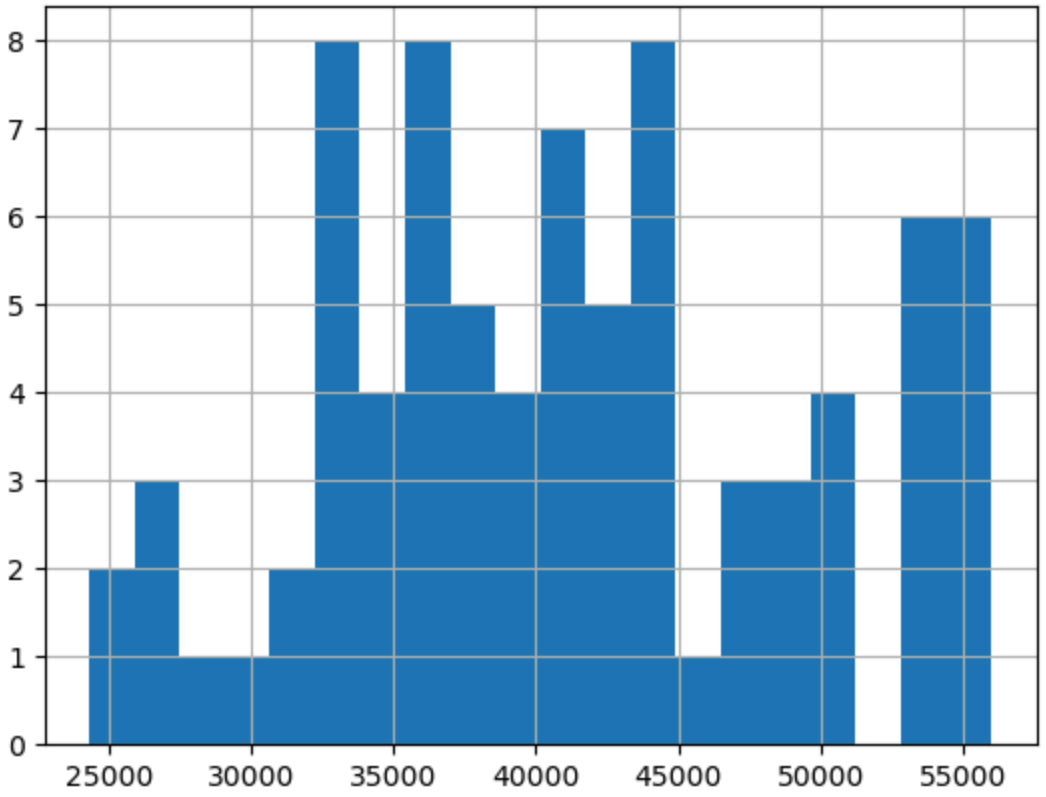
In [3]:
```
## Check unique values and validate if data is clean
cols = df.columns
def Unique_Values():
    for i in np.arange(0,len(cols)):
        print('There are {} of unique values in {} column out of {}'.format(df[cols[i]].nunique(), cols[i], len(df)))
print(Unique_Values())

print('variables with NA values', df.isna().sum())
```

```
There are 7 of unique values in month column out of 81
There are 31 of unique values in day column out of 81
There are 80 of unique values in attend column out of 81
There are 7 of unique values in day_of_week column out of 81
There are 17 of unique values in opponent column out of 81
There are 32 of unique values in temp column out of 81
There are 2 of unique values in skies column out of 81
There are 2 of unique values in day_night column out of 81
There are 2 of unique values in cap column out of 81
There are 2 of unique values in shirt column out of 81
There are 2 of unique values in fireworks column out of 81
There are 2 of unique values in bobblehead column out of 81
None
variables with NA values month          0
day             0
attend          0
day_of_week     0
opponent        0
temp            0
skies           0
day_night       0
cap             0
shirt           0
fireworks       0
bobblehead      0
dtype: int64
```

In [44]:

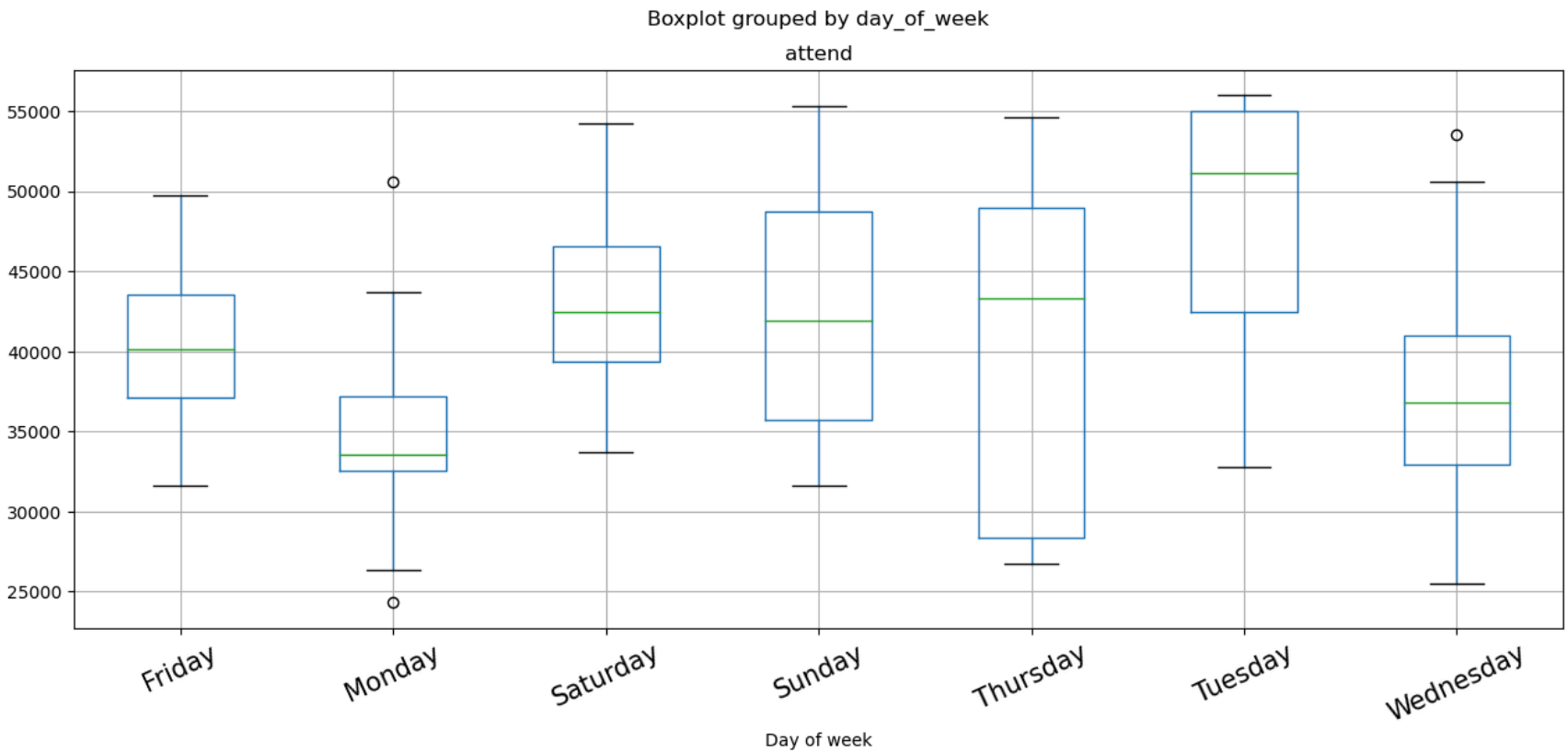
```
# plot a histogram of attendance with a bin size of 20
import matplotlib.pyplot as plt
import statistics

x= statistics.mean(df['attend'])

df['attend'].hist(bins=20)
print ('mean of attendance is:', round(x,0))
```

```
mean of attendance is: 41040.0
```



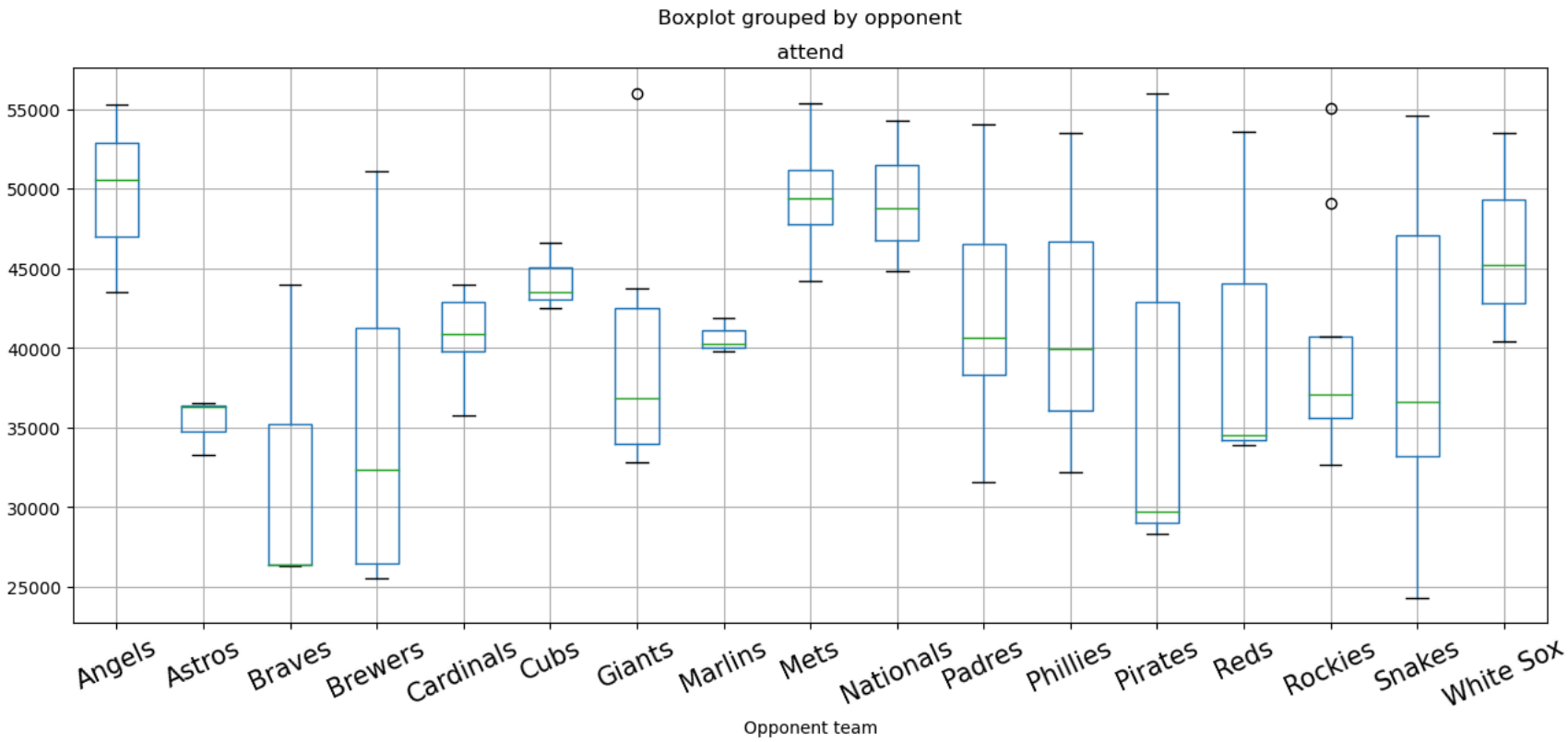
with mean at 41K, we can consider attendance > 45000 is consider as high attendance.

In [50]:
```
#create box plot for attendance vs every variables.

import matplotlib.pyplot as plt

#first, by day of week
df.boxplot(column='attend',by='day_of_week',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("Day of week",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like Tuesday more likely has high attendance, Monday and Wednesday are worst')
```



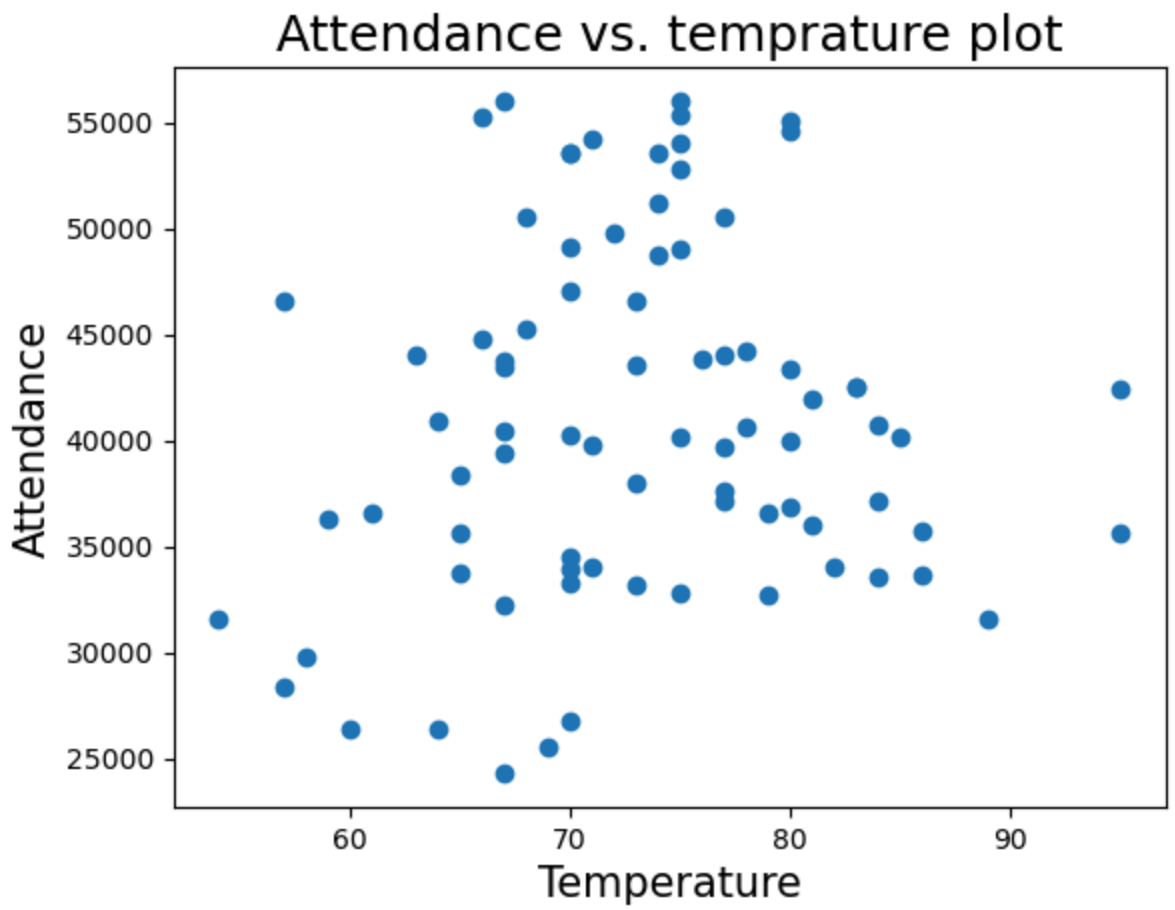
looks like Tuesday more likely has high attendance, Monday and Wednesday are worst

In [42]:
```
#  now by opponents
df.boxplot(column='attend',by='opponent',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("Opponent team",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like games with team Angles, Mets and team Nationals more likely have high attendance')
```



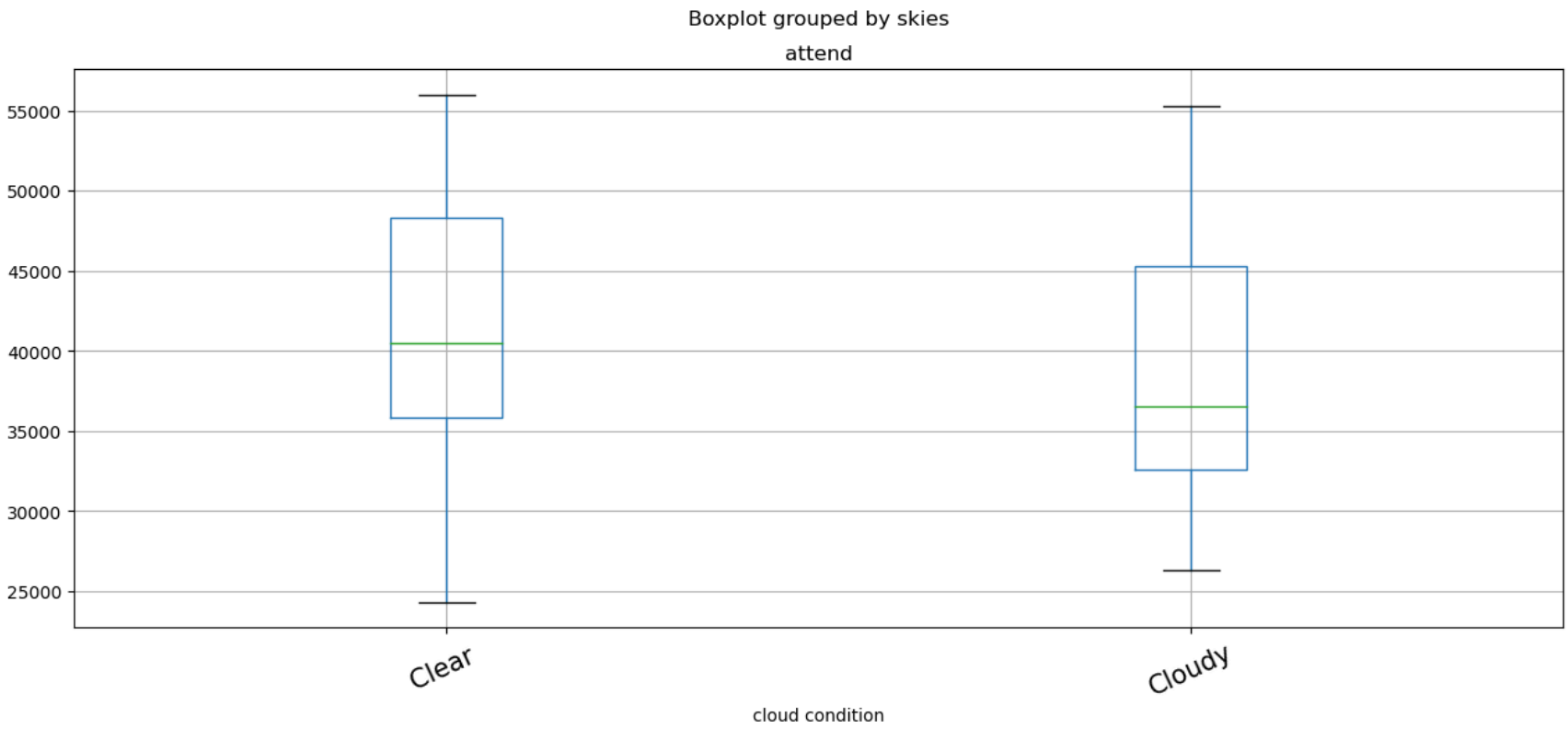
looks like games with team Angles, Mets and team Nationals more likely have high attendance

In [41]:
```
#  now by temps using scatter chart, since temp is a continuous variable

plt.scatter(df['temp'],df['attend'])
plt.title("Attendance vs. temprature plot", fontsize=18)
plt.xlabel("Temperature",fontsize=15)
plt.ylabel("Attendance",fontsize=15)
plt.show()

print('looks like 70-80 degree will more likely attract high attendance,too cold or hot will kill attendance')
```



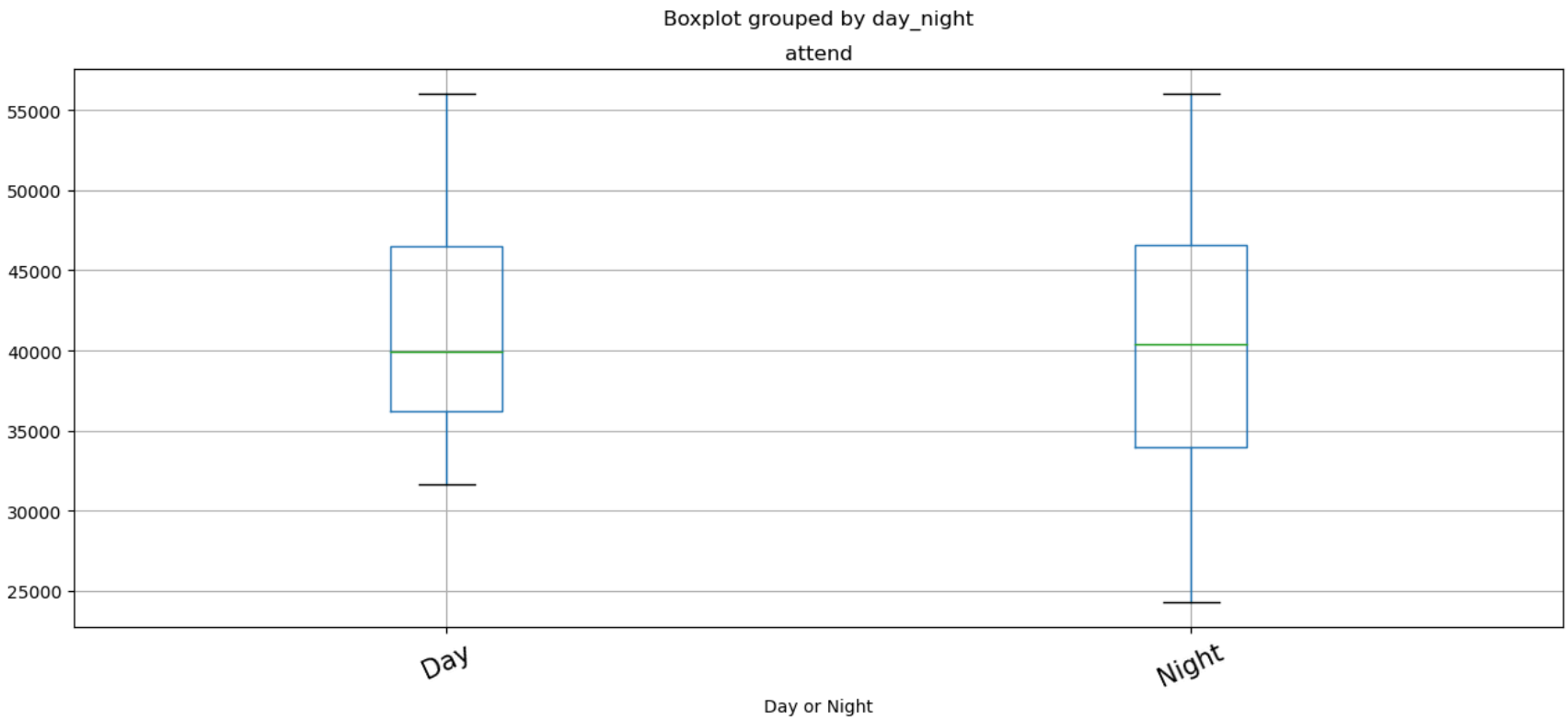
looks like 70-80 degree will more likely attract high attendance,too cold or hot will kill attendance

In [49]:
```
#  now by weather condition
df.boxplot(column='attend',by='skies',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("cloud condition",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('Day with clear sky may have a little more attendance than days with cloudy sky , but not guaranteed')
```



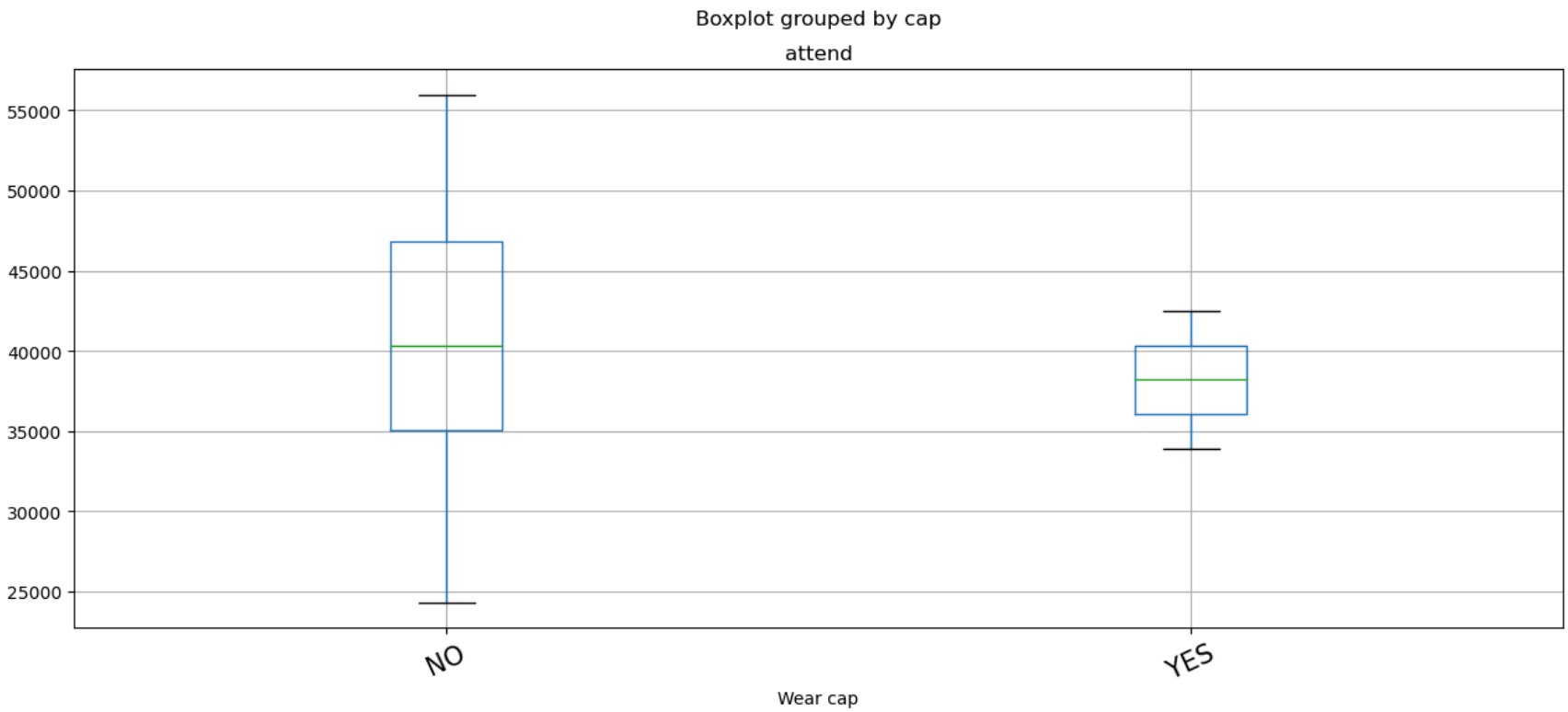
Day with clear sky may have a little more attendance than days with cloudy sky , but not guaranteed

In [36]:
```
#  now by day or night condition
df.boxplot(column='attend',by='day_night',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("Day or Night",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like day or night are also not that important to influence attendance')
```



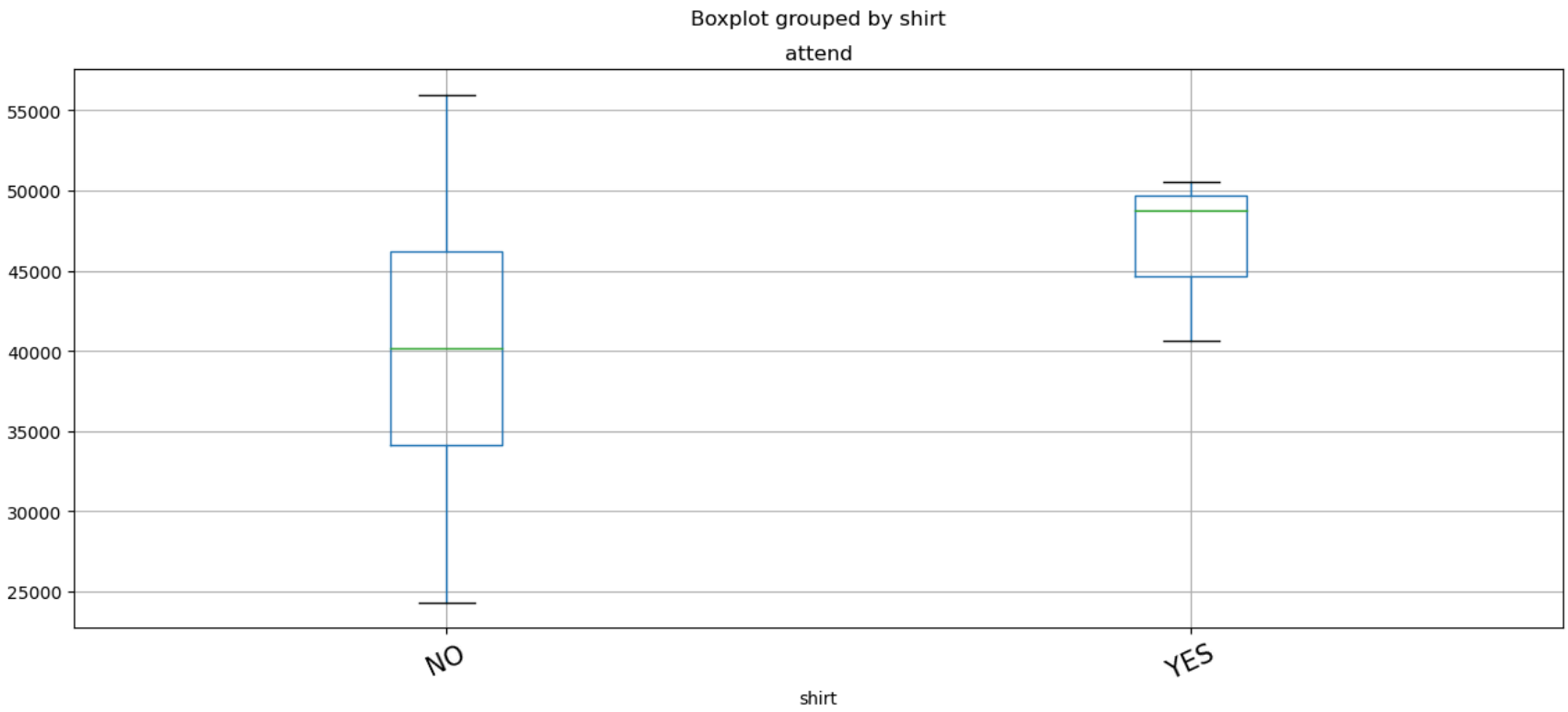
looks like day or night are also not that important to influence attendance

In [40]:
```
#  now by if wear with cap or not condition
df.boxplot(column='attend',by='cap',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("Wear cap",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like cap is also NOT important to influence attendance, but with a cap, attendance is more consistent')
```



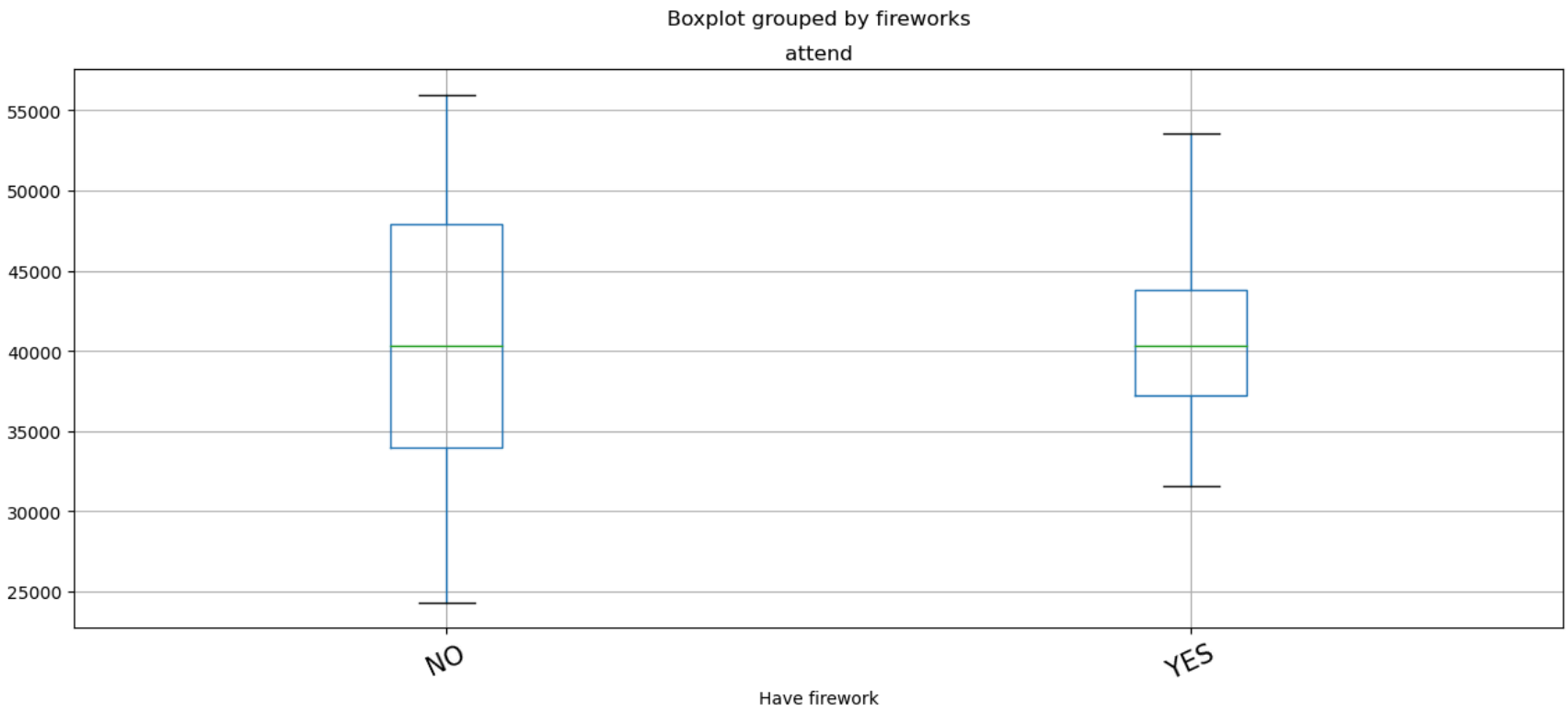
looks like cap is also NOT important to influence attendance, but with a cap, attendance is more consistent

In [28]:
```
#  now check shirt condition

df.boxplot(column='attend',by='shirt',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("shirt",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like wear shirt will improve attendance, also have a more consistent attendance')
```



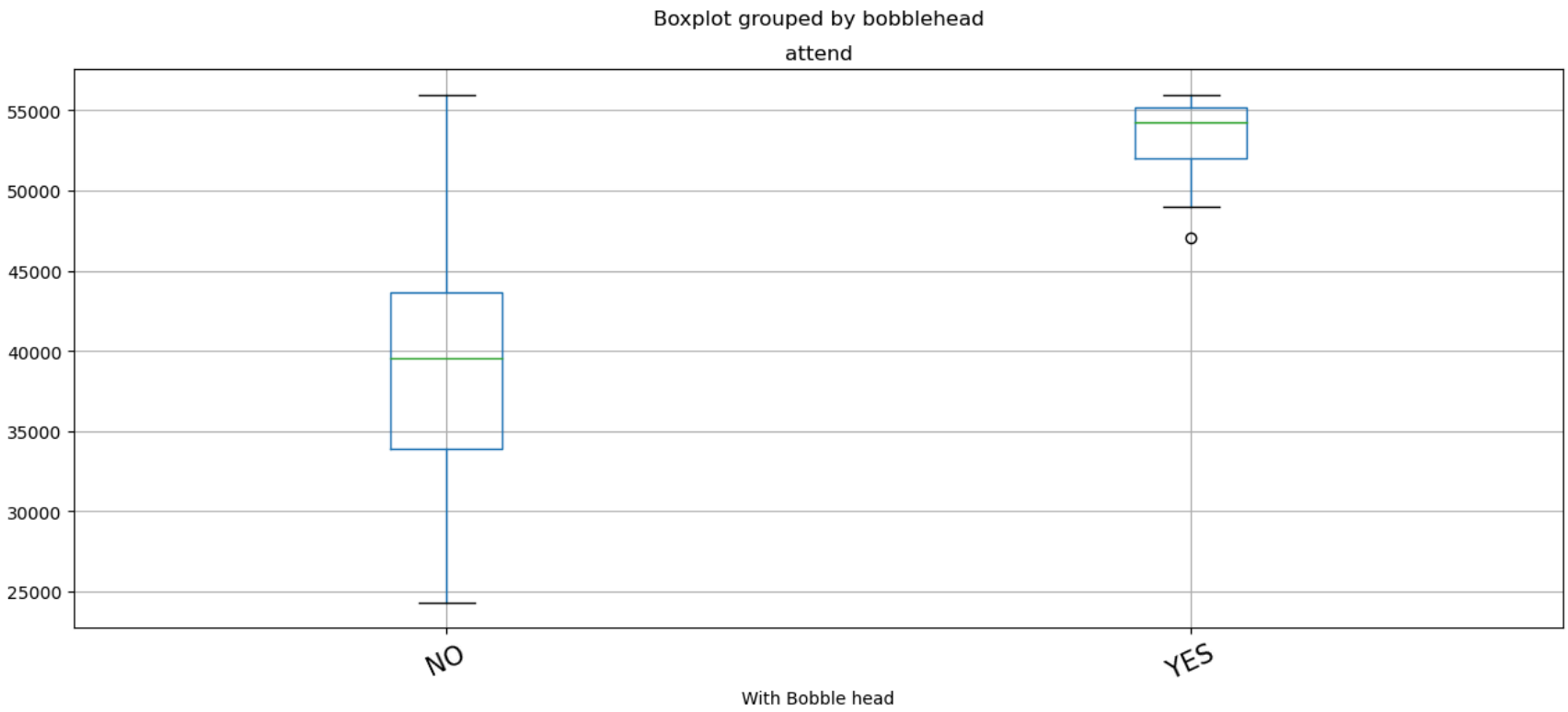
looks like wear shirt will improve attendance, also have a more consistent attendance

In [47]:
```
#  now check firsework

df.boxplot(column='attend',by='fireworks',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("Have firework",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like firwork will not attract high attendance, will bring a more consistent attendance')
```



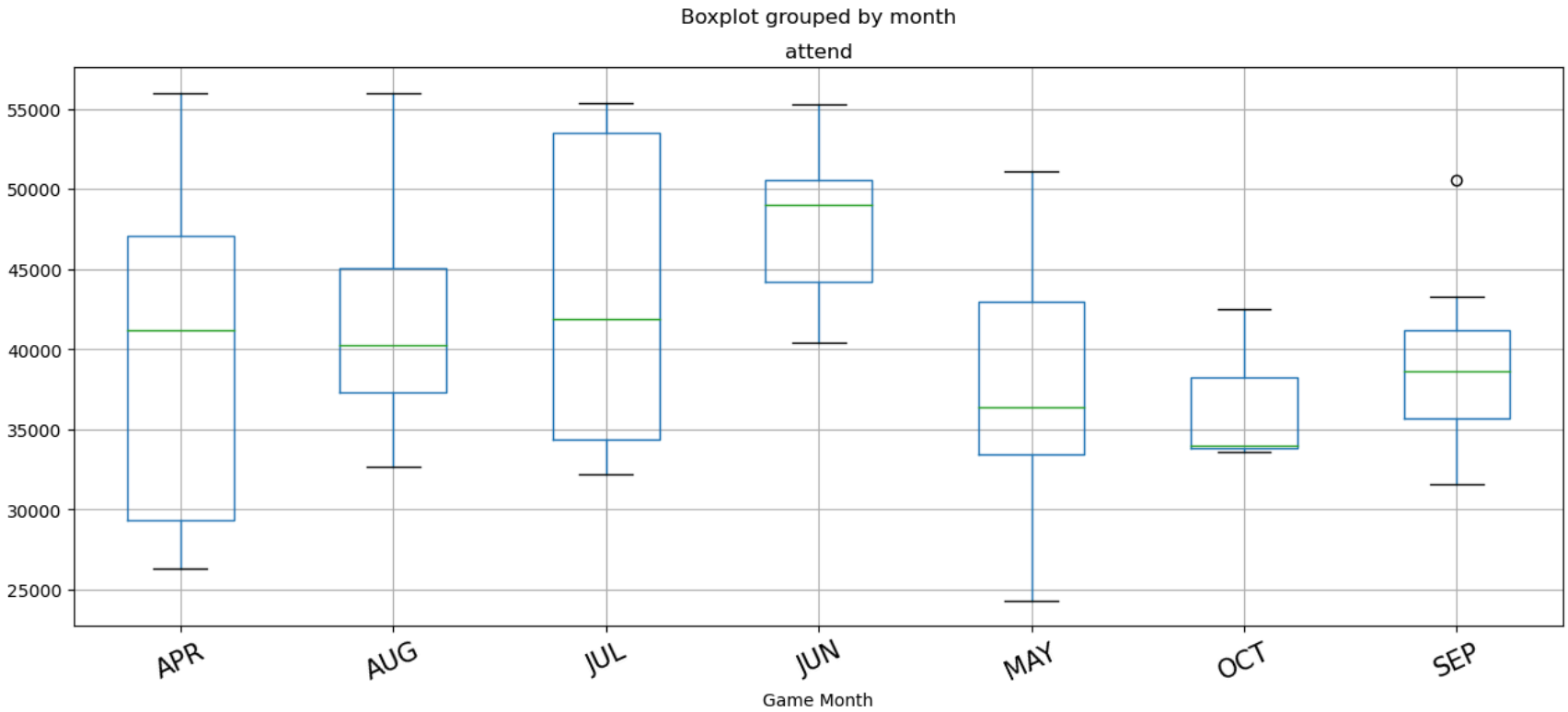
looks like firwork will not attract high attendance, will bring a more consistent attendance

In [48]:
```
#  now check bobble head

df.boxplot(column='attend',by='bobblehead',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("With Bobble head",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like bobble head is important to boost attendance')
```



looks like bobble head is important to boost attendance

In [34]:
```
# Now check month
df.boxplot(column='attend',by='month',figsize=(15,6))
plt.xticks(fontsize=15)
plt.xlabel("Game Month",fontsize=10)
plt.xticks(rotation = 25)
plt.show()
print('looks like warm month like June and July will help attendance')
```




```
looks like warm month like June and July will help attendance
```

Conclusions

1. The factors will not impact attendnace are firwork, cloud condition, if game happens on day or night, as well as wearing cap.
2. A warm temprature, wearing shirt and have bobble head will help to boost up attendance.
3. Games with Angles, Mets and Nationals are likely bring high attendance, games with White Sox is also have some chance to ahieve high attendance.
4. Tuesday is a day very likely have high attendance, Monday and Wednesday are bad days.

Recommendations:

Have games arranged on Tuesday will help to keep attendance stay at a high level(>45K)

have more games on warm month or days. cold (<60F) or hot (>85F) temparture will kill attendance.

Play more games with popular teams, like Mets, Nationals and Angles. people are also interested in games with white sox.

if it is a cold day or other days of week, try to have more booble head, ask team to wear shirt, these actions will help boosting attendance. especially the booble head, it has great power to get high attendance.

Citation: (Bellevue University, 2024) the MLB team dataset: dodgers.csv

In [ ]: