In [1]:

```
# William Barker
# DSC630
# Week 3

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:

```
df = pd.read_csv('dodgers-2022.csv')
df.head()
```

Out[2]:

	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 [3]:

```
# checking for missing values

nan_df = df.isna()
print(nan_df)
```

```
    month    day  attend  day_of_week  opponent   temp  skies  day_night  \
0   False  False   False        False     False  False  False      False
1   False  False   False        False     False  False  False      False
2   False  False   False        False     False  False  False      False
3   False  False   False        False     False  False  False      False
4   False  False   False        False     False  False  False      False
..    ...    ...     ...          ...       ...    ...    ...        ...
76  False  False   False        False     False  False  False      False
77  False  False   False        False     False  False  False      False
78  False  False   False        False     False  False  False      False
79  False  False   False        False     False  False  False      False
80  False  False   False        False     False  False  False      False

      cap  shirt  fireworks  bobblehead
0   False  False      False       False
1   False  False      False       False
2   False  False      False       False
3   False  False      False       False
4   False  False      False       False
..    ...    ...        ...         ...
76  False  False      False       False
77  False  False      False       False
78  False  False      False       False
79  False  False      False       False
80  False  False      False       False

[81 rows x 12 columns]
```

In [4]:

```
# double checking for missing values

df.isnull().values.any()
```

```
False
```

Out[4]:

In [5]:
```
# grouping data by day of the week

grouped = df.groupby('day_of_week')
```

In [7]:
```
# finding the average attendance of each day

average_attendance = grouped['attend'].mean()
print(average_attendance)
```

```
day_of_week
Friday       40116.923077
Monday       34965.666667
Saturday     43072.923077
Sunday       42268.846154
Thursday     40407.400000
Tuesday      47741.230769
Wednesday    37585.166667
Name: attend, dtype: float64
```

In [8]:
```
# grouping data by opponent

grouped1 = df.groupby('opponent')
```

In [10]:
```
# finding average attendance for each opponent

average_attendance1 = grouped1['attend'].mean()
print(average_attendance1)
```

```
opponent
Angels       49777.333333
Astros       35383.333333
Braves       32245.000000
Brewers      35358.750000
Cardinals    40853.285714
Cubs         44206.666667
Giants       39296.333333
Marlins      40665.333333
Mets         49586.250000
Nationals    49267.333333
Padres       42092.222222
Phillies     41897.000000
Pirates      38019.000000
Reds         40649.000000
Rockies      39631.222222
Snakes       39315.444444
White Sox    46382.000000
Name: attend, dtype: float64
```

In [12]:
```
# sorting the data to make it easier to read
# this shows us that on average, tuesdays have the highest attendance, followed by Saturda

sorted_data = average_attendance.sort_values(ascending=False)

sorted_data
```

Out[12]:
```
day_of_week
Tuesday      47741.230769
Saturday     43072.923077
```

```
Sunday       42268.846154
Thursday     40407.400000
Friday       40116.923077
Wednesday    37585.166667
Monday       34965.666667
Name: attend, dtype: float64
```

In [13]:

```
# sorting the data to make it easier to read
# this shows us that games against the Angels, Mets, and Nationals have the highest averag

sorted_data1 = average_attendance1.sort_values(ascending=False)

sorted_data1
```

Out[13]:

```
opponent
Angels       49777.333333
Mets         49586.250000
Nationals    49267.333333
White Sox    46382.000000
Cubs         44206.666667
Padres       42092.222222
Phillies     41897.000000
Cardinals    40853.285714
Marlins      40665.333333
Reds         40649.000000
Rockies      39631.222222
Snakes       39315.444444
Giants       39296.333333
Pirates      38019.000000
Astros       35383.333333
Brewers      35358.750000
Braves       32245.000000
Name: attend, dtype: float64
```

In [14]:

```
# grouping the data by clear or cloudy weather

grouped2 = df.groupby('skies')
```

In [15]:

```
# finding the average attendance based on the skies
# this shows attendance tends to be higher on clear days

average_attendance2 = grouped2['attend'].mean()
print(average_attendance2)
```

```
skies
Clear     41729.209677
Cloudy    38791.315789
Name: attend, dtype: float64
```

In [16]:

```
# grouping the data by whether its day or night

grouped3 = df.groupby('day_night')
```

In [17]:

```
# finding the average attendance based on whether its day or night
# this shows a slight higher average on games during the day

average_attendance3 = grouped3['attend'].mean()
print(average_attendance3)
```

```
day_night
Day      41793.266667
```

```
Night    40868.893939
Name: attend, dtype: float64
```

In [18]:
```
# grouping the data by whether theres fireworks or not

grouped4 = df.groupby('fireworks')
```

In [19]:
```
# finding the average attendance based on whther theres fireworks
# this shows essentially no difference

average_attendance4 = grouped4['attend'].mean()
print(average_attendance4)
```

```
fireworks
NO     41032.179104
YES    41077.857143
Name: attend, dtype: float64
```

In [20]:
```
# grouping the data by bobblehead

grouped5 = df.groupby('bobblehead')
```

In [21]:
```
# finding the average attendance based on whether there is bobblehead or not
# this shows a large increase in average attendance if bobblehead is there

average_attendance5 = grouped5['attend'].mean()
print(average_attendance5)
```

```
bobblehead
NO     39137.928571
YES    53144.636364
Name: attend, dtype: float64
```

In [22]:
```
# sorting our original data frame by attendance
# the two highest attended games were both on Tuesdays, against the Pirates and Giants, bo
# one was at night and the other during the day, neither had fireworks and one had bobbleh

df_sorted = df.sort_values(by='attend', ascending=False)
df_sorted
```

Out[22]:

	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks	bobblehea
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO	N
59	AUG	21	56000	Tuesday	Giants	75	Clear	Night	NO	NO	NO	YE
39	JUL	1	55359	Sunday	Mets	75	Clear	Night	NO	NO	NO	YE
31	JUN	12	55279	Tuesday	Angels	66	Cloudy	Night	NO	NO	NO	YE
56	AUG	7	55024	Tuesday	Rockies	80	Clear	Night	NO	NO	NO	YE
...	...	...	...	...	...	...	...	...	...	...	...	.
29	MAY	31	26773	Thursday	Brewers	70	Clear	Night	NO	NO	NO	N
6	APR	23	26376	Monday	Braves	60	Cloudy	Night	NO	NO	NO	N
8	APR	25	26345	Wednesday	Braves	64	Cloudy	Night	NO	NO	NO	N
28	MAY	30	25509	Wednesday	Brewers	69	Clear	Night	NO	NO	NO	N
18	MAY	14	24312	Monday	Snakes	67	Clear	Night	NO	NO	NO	N

81 rows × 12 columns

In [ ]:

```
# If I were to make recommendations on how to increase attendance, I would suggest having
# often as possible, seeing as "yes" occupies most of the highest attended games and "no"
# attended games. Its impossible to control the weather so there's no point in talking abo
# know much about baseball but if they could play more games against the Mets, Angels or I
# should pull more people. Tuesdays seem to be the most attended games by far (for some re
# always have games on tuesdays and also the weekends.
```