Botts_DSC630_Week3

September 17, 2023

[ ]:
```
# Christopher Botts
# DSC 630
# Week 3
```

[122]:
```
import pandas as pd
import numpy as np
import scipy.stats
import seaborn as sns
from scipy import stats
from scipy.stats import ttest_ind
from matplotlib import pyplot as plt
```

[61]:
```
dodgersdf = pd.read_csv('dodgers-2022.csv')
```

[6]:
```
dodgersdf.shape
```

[6]:
```
(81, 12)
```

[8]:
```
dodgersdf.head()
```

[8]:

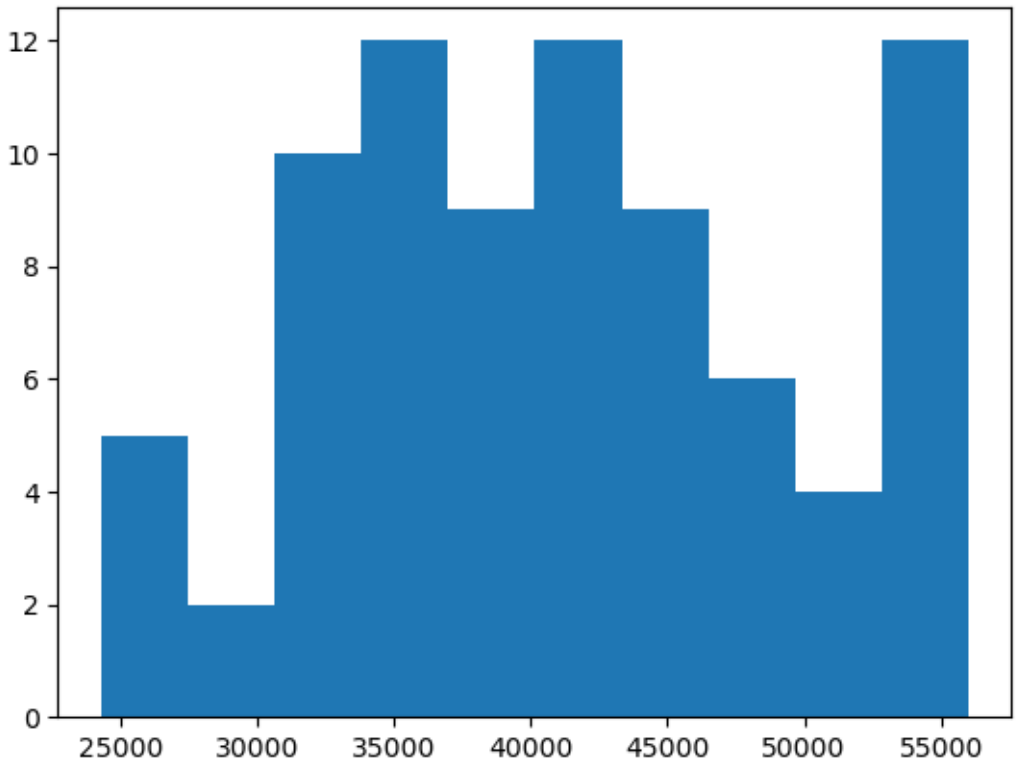
	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks	bobblehead
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO	NO
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	NO	NO	NO	NO
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	NO	NO	NO	NO
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	NO	NO	YES	NO
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	NO	NO	NO	NO

[14]:
```
dodgersdf.dtypes
```

[14]:
```
month          object
day             int64
```

```
attend          int64
day_of_week    object
opponent       object
temp            int64
skies          object
day_night      object
cap            object
shirt          object
fireworks      object
bobblehead     object
dtype: object
```

[11]:
```
plt.hist(dodgersdf['attend'])
plt.show()
```



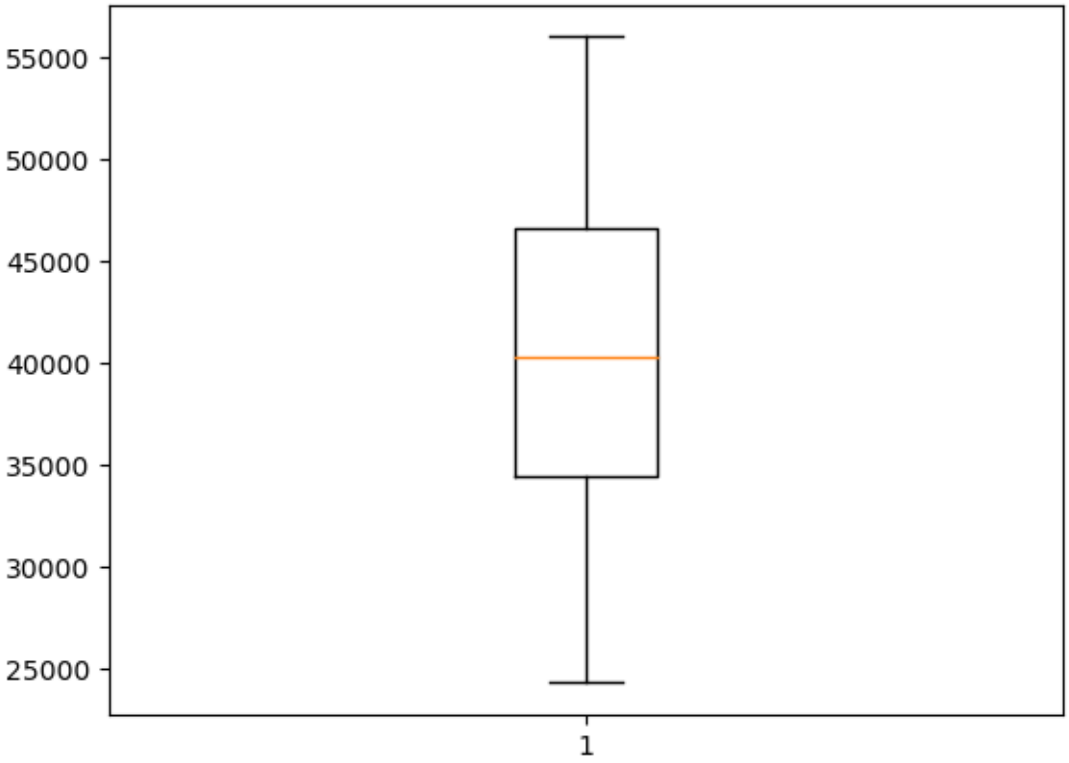
[158]:
```
# calculate summary statistics for the target variable
dodgersdf['attend'].describe()
```

[158]:
```
count       81.000000
mean     41040.074074
std       8297.539460
min      24312.000000
```

```
25%      34493.000000
50%      40284.000000
75%      46588.000000
max      56000.000000
Name: attend, dtype: float64
```

[129]:
```
# create a boxplot to visualize the distribution of the data
plt.boxplot(dodgersdf['attend'])
```

[129]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x173a723e0>,
  <matplotlib.lines.Line2D at 0x173a72680>],
 'caps': [<matplotlib.lines.Line2D at 0x173a72920>,
  <matplotlib.lines.Line2D at 0x173a72bc0>],
 'boxes': [<matplotlib.lines.Line2D at 0x173a72140>],
 'medians': [<matplotlib.lines.Line2D at 0x173a72e60>],
 'fliers': [<matplotlib.lines.Line2D at 0x173a73100>],
 'means': []}
```



The boxplot shows a symmetrical distribution of the data for attendance numbers.

[137]:
```
# use the z score to identify outliers
z = np.abs(stats.zscore(dodgersdf['attend']))
```

```
z
```

[137]:
```
0     1.814169
1     1.371678
2     1.541575
3     1.144663
4     0.668059
        …
76    0.038330
77    0.658861
78    0.899337
79    0.173769
80    0.852042
Name: attend, Length: 81, dtype: float64
```

[139]:
```
print(np.where(z > 2))

print(z[18])
```

```
(array([18]),)
2.0285893093559086
```

There is one outlier from the attendance attribute

[152]:
```
# extract information about the game that is an outlier for attendance.
dodgersdf.iloc[18]
```

[152]:
```
month            MAY
day               14
attend         24312
day_of_week    Monday
opponent       Snakes
temp               67
skies          Clear
day_night       Night
cap                NO
shirt              NO
fireworks          NO
bobblehead         NO
Name: 18, dtype: object
```

The outlier game has an attendance number that is below 2 times the standard deviation from the mean. However, when viewing the details of that game, there is not immediate explanation why attendance was significantly low on this day.

1 Evaluate attendance by the time of day

[164]:
```
# Compare the attendance of games by the time of day the game was held
print("There were", len(dodgersdf[dodgersdf['day_night'] == 'Day']), "games␣
 ↪held during the daytime.\n")
print("There were", len(dodgersdf[dodgersdf['day_night'] == 'Night']), "games␣
 ↪held during the nighttime.")
```

```
There were 15 games held during the daytime.

There were 66 games held during the nighttime.
```

[168]:
```
#calculate the average attendace for nighttime games
nightgames = dodgersdf[dodgersdf['day_night'] == 'Night']
nightgames['attend'].sum()/len(nightgames)
```

[168]:
```
40868.893939393936
```

[169]:
```
# calculate the average attendance for daytime games
daygames = dodgersdf[dodgersdf['day_night'] == 'Day']
daygames['attend'].sum()/len(daygames)
```

[169]:
```
41793.26666666667
```

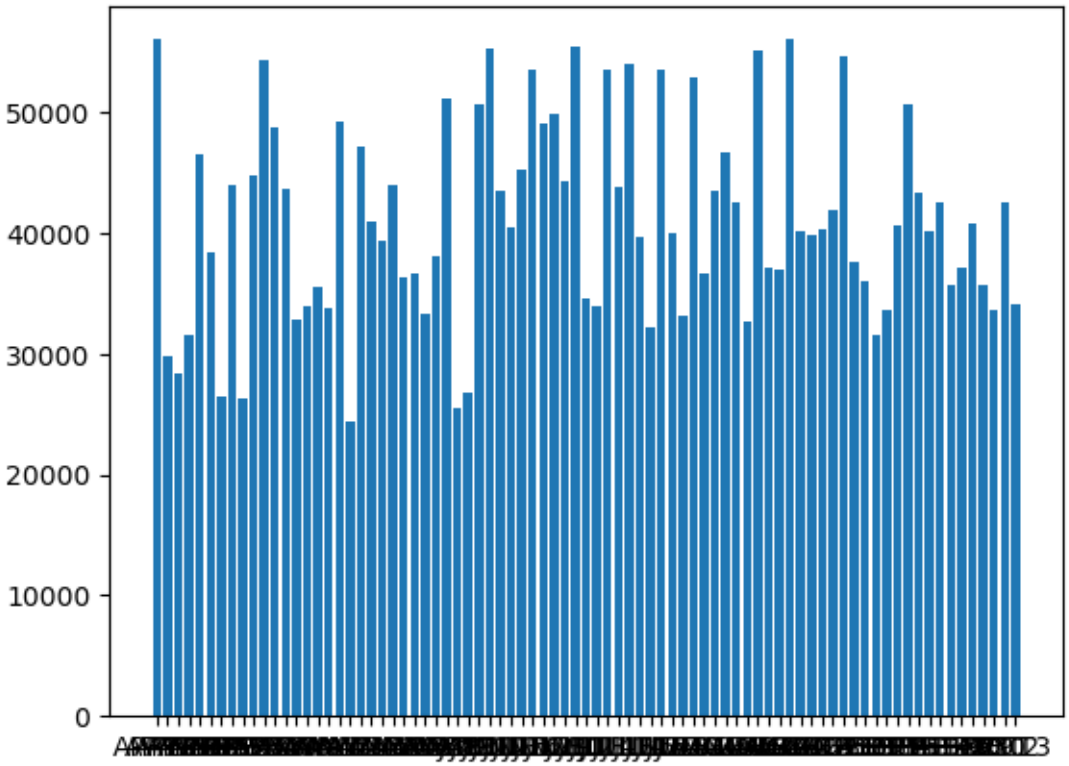
There is a similar average attendance for daytime and nighttime games.

2 Plot attendance numbers over the course of the season

[24]:
```
# create a feature that displays the date of the game
dodgersdf['date'] = dodgersdf['month'] + dodgersdf['day'].astype(str)
dodgersdf['date']
```

[26]:
```
# plot the attendance numbers by day
plt.bar(dodgersdf['date'], height=dodgersdf['attend'])
```

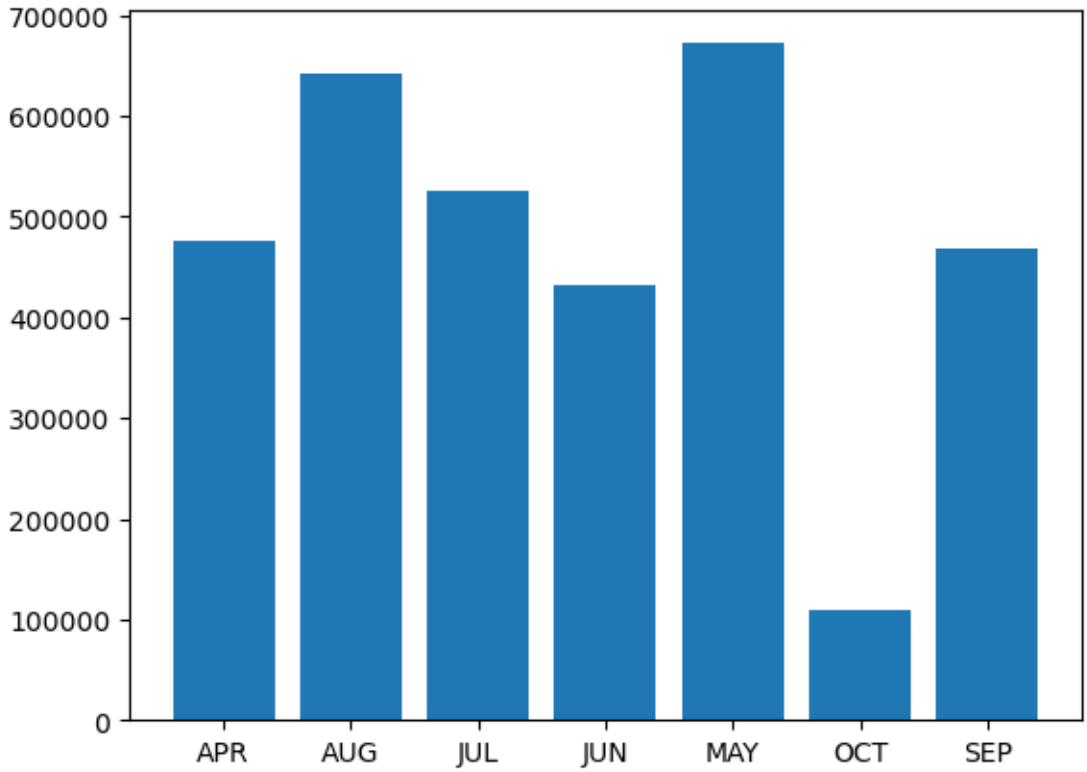
[26]:
```
<BarContainer object of 81 artists>
```



[48]:
```
# plot the attendance numbers by month
monthly = dodgersdf['attend'].groupby(dodgersdf['month']).sum()

plt.bar(x=monthly.index, height=monthly)
```

[48]: <BarContainer object of 7 artists>



3 Look for correlations in the dataset features

[63]:
```
# one hot encode the dataset
dodgersdum = pd.get_dummies(dodgersdf)
```

[64]:
```
dodgersdum.shape
```

[64]:
```
(81, 46)
```

[126]:
```
dodgersdum.columns
```

[126]:
```
Index(['day', 'attend', 'temp', 'month_APR', 'month_AUG', 'month_JUL',
       'month_JUN', 'month_MAY', 'month_OCT', 'month_SEP',
       'day_of_week_Friday', 'day_of_week_Monday', 'day_of_week_Saturday',
       'day_of_week_Sunday', 'day_of_week_Thursday', 'day_of_week_Tuesday',
       'day_of_week_Wednesday', 'opponent_Angels', 'opponent_Astros',
       'opponent_Braves', 'opponent_Brewers', 'opponent_Cardinals',
       'opponent_Cubs', 'opponent_Giants', 'opponent_Marlins', 'opponent_Mets',
       'opponent_Nationals', 'opponent_Padres', 'opponent_Phillies',
       'opponent_Pirates', 'opponent_Reds', 'opponent_Rockies',
       'opponent_Snakes', 'opponent_White Sox', 'skies_Clear ', 'skies_Cloudy',
```

```
       'day_night_Day', 'day_night_Night', 'cap_NO', 'cap_YES', 'shirt_NO',
       'shirt_YES', 'fireworks_NO', 'fireworks_YES', 'bobblehead_NO',
       'bobblehead_YES'],
      dtype='object')
```

[71]:
```
# create a correlation matrix
corr_df = dodgersdum.corr()
```

[97]:
```
corr_df['attend']
```

[97]:
```
day                      0.027093
attend                   1.000000
temp                     0.098951
month_APR               -0.073237
month_AUG                0.098944
month_JUL                0.143837
month_JUN                0.295853
month_MAY               -0.239471
month_OCT               -0.103132
month_SEP               -0.105443
day_of_week_Friday      -0.048948
day_of_week_Monday      -0.307198
day_of_week_Saturday     0.107788
day_of_week_Sunday       0.065153
day_of_week_Thursday    -0.019679
day_of_week_Tuesday      0.355316
day_of_week_Wednesday   -0.174723
opponent_Angels          0.207796
opponent_Astros         -0.134533
opponent_Braves         -0.209171
opponent_Brewers        -0.157030
opponent_Cardinals      -0.006967
opponent_Cubs            0.075310
opponent_Giants         -0.074763
opponent_Marlins        -0.008912
opponent_Mets            0.236213
opponent_Nationals       0.195667
opponent_Padres          0.045111
opponent_Phillies        0.020380
opponent_Pirates        -0.071849
opponent_Reds           -0.009301
opponent_Rockies        -0.060404
opponent_Snakes         -0.073943
opponent_White Sox       0.127046
skies_Clear              0.150963
skies_Cloudy            -0.150963
day_night_Day            0.043544
```

```
day_night_Night         -0.043544
cap_NO                   0.055002
cap_YES                 -0.055002
shirt_NO                -0.133269
shirt_YES                0.133269
fireworks_NO            -0.002094
fireworks_YES            0.002094
bobblehead_NO           -0.581895
bobblehead_YES           0.581895
Name: attend, dtype: float64
```

The dataset was one hot encoded so that correlation of attendance and categorical features could be analyzed. The datasets shows mostly weak and moderate correlations. The stronger correlations should be further evaluated.

4 Analzye the effect of opponent on game attendance

[117]:
```
# calculate the total number of attendees of games grouped by the opponent␣
↪played
dodgersdf['attend'].groupby(dodgersdf['opponent']).sum()
```

[117]:
```
opponent
Angels       149332
Astros       106150
Braves        96735
Brewers      141435
Cardinals    285973
Cubs         132620
Giants       353667
Marlins      121996
Mets         198345
Nationals    147802
Padres       378830
Phillies     125691
Pirates      114057
Reds         121947
Rockies      356681
Snakes       353839
White Sox    139146
Name: attend, dtype: int64
```

[124]:
```
# count the number of games each opponent played
x = dodgersdf['attend'].groupby(dodgersdf['opponent']).count()
print(x)
```

```
opponent
Angels       3
```

```
Astros       3
Braves       3
Brewers      4
Cardinals    7
Cubs         3
Giants       9
Marlins      3
Mets         4
Nationals    3
Padres       9
Phillies     3
Pirates      3
Reds         3
Rockies      9
Snakes       9
White Sox    3
Name: attend, dtype: int64
```

[125]:
```
# calculate the average attendance of dodgers game by opponent played
y = dodgersdf['attend'].groupby(dodgersdf['opponent']).sum() /␣
  ↪dodgersdf['attend'].groupby(dodgersdf['opponent']).count()
print(y)
```

```
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
```

[123]:
```
# check the significance of the correlation between the number of games played␣
  ↪against a specific opponent and the average number of attendees
scipy.stats.pearsonr(x, y)
```

[123]:
```
PearsonRResult(statistic=-0.15576831495849688, pvalue=0.5505126763690786)
```

This calculation was made to determine if the number of games played against a particular opponent influenced the number of attendees. There is a slightly negative correlation, indicating that the less number of times a particular opponent was played, the more fans attended the game. However, this correlation is not statistically signifcant when evaluated with Pearson's R and a .05 significance value.

[206]:
```
# check the significance of the strongest corrlated opponent, the Mets.
scipy.stats.pearsonr(dodgersdum['attend'], dodgersdum['opponent_Mets'])
```

[206]:
```
PearsonRResult(statistic=0.23621346551829403, pvalue=0.03375208420005861)
```

The strongest correlation with opponents occurred between attendance the Mets. When this relationship was evaluated on its own, there is a statistically significant relationship.

5 Build a list of statistically significant correlations

[175]:
```
# create a list of statistically significant correlations with attendance
attrib = [x for x in dodgersdum.columns]
attrib
```

[175]:
```
['day',
 'attend',
 'temp',
 'month_APR',
 'month_AUG',
 'month_JUL',
 'month_JUN',
 'month_MAY',
 'month_OCT',
 'month_SEP',
 'day_of_week_Friday',
 'day_of_week_Monday',
 'day_of_week_Saturday',
 'day_of_week_Sunday',
 'day_of_week_Thursday',
 'day_of_week_Tuesday',
 'day_of_week_Wednesday',
 'opponent_Angels',
 'opponent_Astros',
 'opponent_Braves',
 'opponent_Brewers',
 'opponent_Cardinals',
 'opponent_Cubs',
 'opponent_Giants',
 'opponent_Marlins',
 'opponent_Mets',
 'opponent_Nationals',
```

```
 'opponent_Padres',
 'opponent_Phillies',
 'opponent_Pirates',
 'opponent_Reds',
 'opponent_Rockies',
 'opponent_Snakes',
 'opponent_White Sox',
 'skies_Clear ',
 'skies_Cloudy',
 'day_night_Day',
 'day_night_Night',
 'cap_NO',
 'cap_YES',
 'shirt_NO',
 'shirt_YES',
 'fireworks_NO',
 'fireworks_YES',
 'bobblehead_NO',
 'bobblehead_YES']
```

[203]:
```
corr_lst = []

for x in attrib:
    y = (scipy.stats.pearsonr(dodgersdum['attend'], dodgersdum[x]))
    c = (x, y[0], y[1])

    if y[0] == 1.0:
        pass
    elif y[1] < .05:
        corr_lst.append(c)
    else:
        pass
```

[204]:
```
corr_lst
```

[204]:
```
[('month_JUN', 0.2958527412896723, 0.007327007423949007),
 ('month_MAY', -0.23947072157291688, 0.031305112591933935),
 ('day_of_week_Monday', -0.30719785832757923, 0.005277280671934742),
 ('day_of_week_Tuesday', 0.3553163421794233, 0.0011337033503210417),
 ('opponent_Mets', 0.23621346551829403, 0.03375208420005861),
 ('bobblehead_NO', -0.5818949681431957, 1.2169642509120652e-08),
 ('bobblehead_YES', 0.5818949681431956, 1.216964250912072e-08)]
```

Pearson's coeffecient shows a statistically significant correlation between attendance and the month of May, the month of June, Monday games, Tuesday games, games against the Mets, and bobbleheads.

6 Analyze correlation between monthly attendees

There is a statistically significant positive correlation between games that occur in June and the number of game attendees and a significant negative correlation between games that occur in May. It seems plausable that the more games that are played in a month, the less attendees would be present, so it should be evaluated whether this correlation exists because of the number of games played in the month.

[223]:
```
# count the number of games played by month
dodgersdf['attend'].groupby(dodgersdf['month']).count()
```

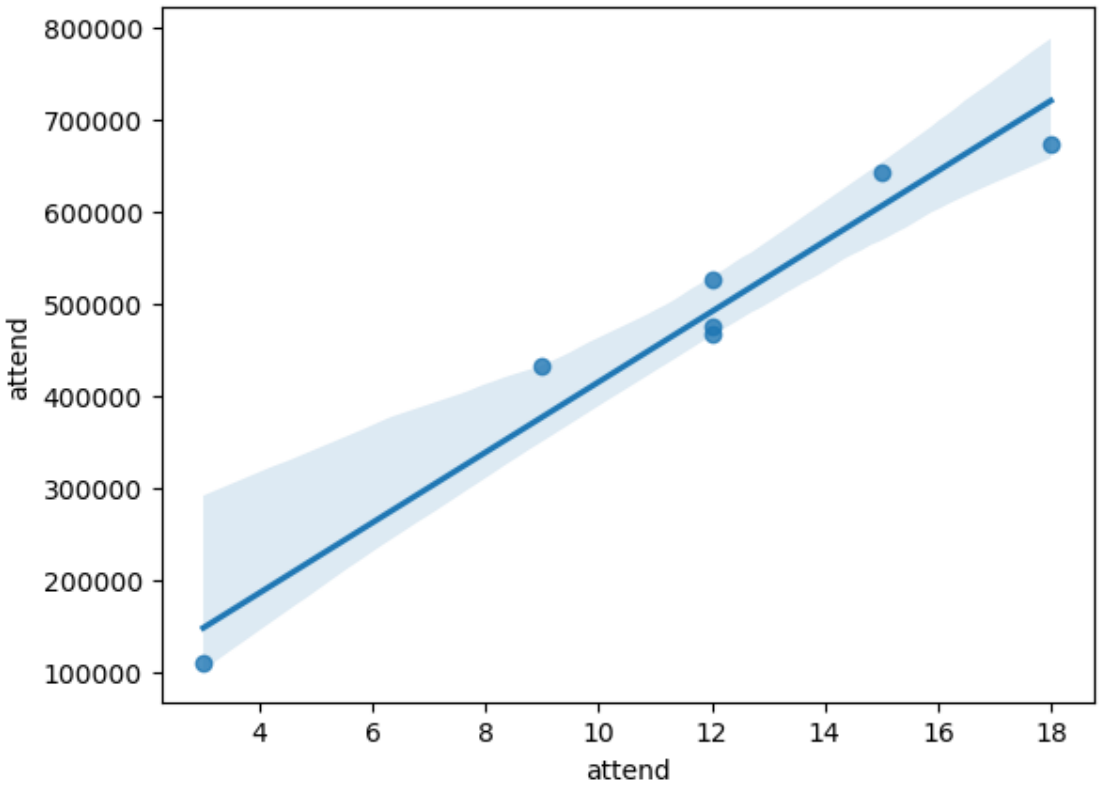
[223]:
```
month
APR    12
AUG    15
JUL    12
JUN     9
MAY    18
OCT     3
SEP    12
Name: attend, dtype: int64
```

[209]:
```
# calculate the average number of attendees per game by month
dodgersdf['attend'].groupby(dodgersdf['month']).sum() / dodgersdf['attend'].
 ↪groupby(dodgersdf['month']).count()
```

[209]:
```
month
APR    39591.916667
AUG    42751.533333
JUL    43884.250000
JUN    47940.444444
MAY    37345.722222
OCT    36703.666667
SEP    38955.083333
Name: attend, dtype: float64
```

[227]:
```
# plot the number of attendees against the number of games played in a month
sns.regplot(data=dodgersdf, x=dodgersdf['attend'].groupby(dodgersdf['month']).
 ↪count(), y=dodgersdf['attend'].groupby(dodgersdf['month']).sum())
```

[227]:
```
<Axes: xlabel='attend', ylabel='attend'>
```



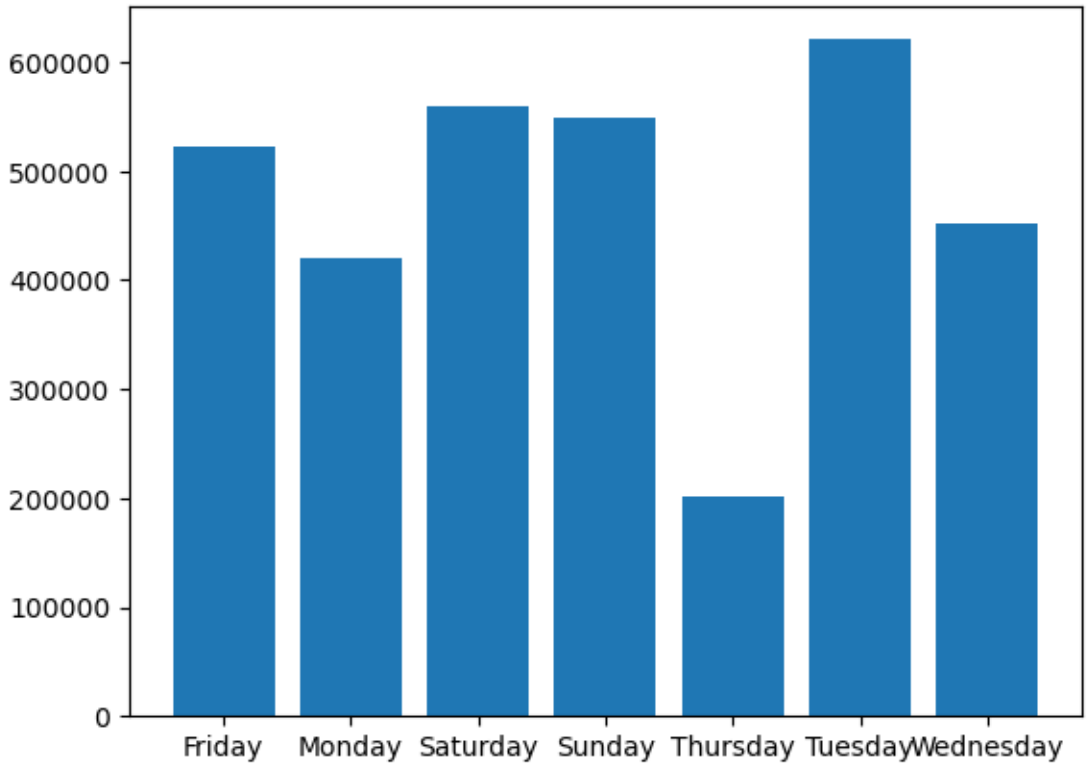
This scatter plots shows the number of total attendees to dodgers games per month and the number of games played per month. The line of best fit suggests that the number of games played in a month does not affect the average number of attendees. Therefore, the correlation observed during the month of June and May are likely due to another reason.

7 Analyze the number of attendees by day of the week

[240]:
```
# plot the attendance numbers by day of the week
monthly = dodgersdf['attend'].groupby(dodgersdf['day_of_week']).sum()

plt.bar(x=monthly.index, height=monthly)
```

[240]:
```
<BarContainer object of 7 artists>
```



There is a positive correlation between Tuesday and attendance and a negative correlation between Monday and attendance. Tuesday is also the day of the week that has the most attendees, while Monday has a relatively small number of attendees.

8 Analyze the bobblehead correlation

[239]:
```
# use a t test to determine if bobblehead is a significant predictor of␣
↪attendance.
sample1 = dodgersdf[dodgersdf['bobblehead']=='YES']
sample2 = dodgersdf[dodgersdf['bobblehead']=='NO']

ttest_ind(sample1['attend'], sample2['attend'])
```

[239]:
```
Ttest_indResult(statistic=6.359553539813022, pvalue=1.2169642509120423e-08)
```

It is unclear to what the variable 'bobblehead' refers. However, it has a strong correlation with dodger game attendance (.582). This correlation is statistically significant when using a t test as a comparison of means.

9 Conclusion

There appears to be several opportunities to increase attendance at Dodger games. Some of these opportunites relate to specific days and times of the year during which a game is played. The month of June has a strong positive correlation with attendance, as well as the highest average attendance per game. I would recommend reallocating marketing resources from June to less popular months, such as May and October. May exhibits a negative correlation with attendance. To address this, I would market Tuesday games more heavily during the month of May. The reason for this is because the positive correlation with attendance and Tuesdays suggests that baseball fans prefer to go to Tuesday games, therefore, marketers will likely have the most success increasing attendance on Tuesdays in May. Thursdays have the lowest attendance. I would also recommend developing a promotion for Thursday games to entice fans to attend these games. Certain opponents also are linked with an increase in attendance: specifically the Mets. Games played against the Mets should be more heavily marketed because there is likely to be a good ROI on these marketing investments. Lastly, the bobblehead correlation should be exploited. Bobblehead is the feature with the strongest correlation with attendance. Furthermore, our signifcance test indicate we can have a strong level of confidence in this correlation.