Alex Spence

DSC 630 Week 3

Assignment 3.2: Using Data to Improve MLB Attendance

In this assignment, you will be using data on the Los Angeles Dodgers Major League Baseball (MLB) team located here: dodgers.csv. Use this data to make a recommendation to management on how to improve attendance. Tell a story with your analysis and clearly explain the steps you take to arrive at your conclusion. This is an open-ended question, and there is no one right answer. You are welcome to do additional research and/or use domain knowledge to assist your analysis, but clearly state any assumptions you make.

Step 1: Loading and Exploring the Data

In [9]:

```
import statsmodels.api as sm
import pandas as pd

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

# Clean minor issues: convert numeric columns and strip spaces
df['attend'] = pd.to_numeric(df['attend'])
df['temp'] = pd.to_numeric(df['temp'])
df['skies'] = df['skies'].str.strip()

# Compute summary statistics
attendance_summary = df['attend'].describe()

# Print results
print("Attendance Summary Statistics:")
print(attendance_summary)
```

```
Attendance Summary Statistics:
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
```

Load the data from the CSV file into a Pandas DataFrame and perform basic cleaning. Ensuring numeric types for attendance and temperature, stripping extra spaces from categorical fields like 'skies'.

The summary statistics provide an overview: 81 games, mean attendance of 41,040, standard deviation of 8,298, minimum of 24,312, and maximum of 56,000 (sellout).

This indicates strong overall attendance (about 73% of stadium capacity on average) but opportunities to boost lower-attendance games.

The median (40,284) is close to the mean, suggesting a symmetric distribution without extreme skew.

Step 2: Grouping and Comparing Key Factors

In [3]:

```
# Group by day of the week
day_of_week_avg = df.groupby('day_of_week')['attend'].mean().sort_values(asc

# Group by month
month_avg = df.groupby('month')['attend'].mean().sort_values(ascending=False

# Group by opponent
opponent_avg = df.groupby('opponent')['attend'].mean().sort_values(ascending

# Day vs. Night
day_night_avg = df.groupby('day_night')['attend'].mean()

# Skies
skies_avg = df.groupby('skies')['attend'].mean()

# Temperature correlation
temp_corr = df['attend'].corr(df['temp'])

# Print results
print("\nAverage Attendance by Day of Week:")
print(day_of_week_avg)
print("\nAverage Attendance by Month:")
print(month_avg)
print("\nAverage Attendance by Opponent:")
print(opponent_avg)
print("\nAverage Attendance by Day/Night:")
print(day_night_avg)
print("\nAverage Attendance by Skies:")
print(skies_avg)
print(f"\nCorrelation between Temperature and Attendance: {temp_corr:.3f}")
```

```
Average Attendance by Day of Week:
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

Average Attendance by Month:
month
JUN    47940.444444
JUL    43884.250000
AUG    42751.533333
APR    39591.916667
SEP    38955.083333
MAY    37345.722222
OCT    36703.666667
Name: attend, dtype: float64

Average Attendance by Opponent:
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

Average Attendance by Day/Night:
day_night
Day      41793.266667
Night    40868.893939
Name: attend, dtype: float64

Average Attendance by Skies:
skies
Clear     41729.209677
Cloudy    38791.315789
Name: attend, dtype: float64

Correlation between Temperature and Attendance: 0.099
```

Grouped attendance by key categorical factors to identify patterns.

Tuesdays (47,741) and weekends outperform weekdays, possibly due to promotions or scheduling.

Summer months (June: 47,940) draw more fans than spring (April: 39,592) or fall, aligning with vacation periods and warmer weather.

Rival games (e.g., Angels: 49,777) boost crowds compared to others (e.g., Braves: 32,245).

Day games (41,793) slightly edge night games, and clear skies (41,729) beat cloudy (38,791).

Temperature shows a weak positive correlation (r=0.099), meaning it's not a major driver in LA's mild climate.

These groupings highlight factors influencing turnout.

Step 3: Analyzing Promotions

In [4]:

```
# List of promotion columns
promos = ['cap', 'shirt', 'fireworks', 'bobblehead']

# Dictionary to store averages and counts
promo_analysis = {}

for promo in promos:
    # Average attendance with/without promo
    avg_yes = df[df[promo] == 'YES']['attend'].mean()
    avg_no = df[df[promo] == 'NO']['attend'].mean()
    count_yes = df[df[promo] == 'YES'].shape[0]
    promo_analysis[promo] = {
        'YES': {'avg': avg_yes, 'count': count_yes},
        'NO': {'avg': avg_no}
    }

# Print results
print("\nPromotion Analysis:")
for promo, data in promo_analysis.items():
    print(f"{promo.capitalize()}: YES (Avg: {data['YES']['avg']:.0f}, Count:
```

```
Promotion Analysis:
Cap: YES (Avg: 38190, Count: 2) | NO (Avg: 41112)
Shirt: YES (Avg: 46644, Count: 3) | NO (Avg: 40825)
Fireworks: YES (Avg: 41078, Count: 14) | NO (Avg: 41032)
Bobblehead: YES (Avg: 53145, Count: 11) | NO (Avg: 39138)
```

I decided to focus on promotions here since management can directly influence this.

Bobbleheads show the strongest boost (+14,007 fans on average, used in 11 games), pushing attendance near sellouts.

Shirts (+5,819, 3 games) provide a modest lift, while fireworks (negligible difference, 14 games) and caps (-2,922, 2 games) show mixed or no clear impact.

The fireworks effect may be diluted as they're often on Fridays which already have high attendance. The small sample sizes for some promotions limit conclusions, but bobbleheads seem to really be driving attendance increase.

Step 4: Regression Analysis for Deeper Insights

In [12]:

```
import statsmodels.api as sm

# Prepare data for regression: convert promotions to binary (1/0)
df_reg = df.copy()
for promo in promos:
    df_reg[promo] = (df_reg[promo] == 'YES').astype(int)

# Create dummy variables for categoricals
df_reg = pd.get_dummies(df_reg, columns=['month', 'day_of_week', 'opponent',

# Drop non-predictor columns (e.g., day is redundant)
df_reg = df_reg.drop(['day'], axis=1)

# Define X (predictors) and y (target)
X = df_reg.drop('attend', axis=1)
X = sm.add_constant(X)  # Add intercept
X = X.astype(float)  # Ensure all are numeric
y = df_reg['attend']

# Fit OLS model
model = sm.OLS(y, X).fit()

# Print summary
print("\nRegression Summary:")
print(model.summary())
```

```
Regression Summary:
                            OLS Regression Results
=============================================================================
==
Dep. Variable:                 attend   R-squared:                       0.7
09
Model:                            OLS   Adj. R-squared:                  0.4
82
Method:                 Least Squares   F-statistic:                     3.1
27
Date:                Tue, 23 Sep 2025   Prob (F-statistic):           0.0001
87
Time:                        21:01:21   Log-Likelihood:                 -795.
41
No. Observations:                  81   AIC:                             166
3.
Df Residuals:                      45   BIC:                             174
9.
Df Model:                          35
Covariance Type:            nonrobust
=============================================================================
=============
                            coef    std err          t      P>|t|      [0.02
5      0.975]
-----------------------------------------------------------------------------
-------------
const                  2.577e+04   1.89e+04      1.361      0.180   -1.24e+0
4    6.39e+04
temp                     11.6874    245.630      0.048      0.962    -483.03
6     506.411
cap                   -6432.7486   5865.472     -1.097      0.279   -1.82e+0
4    5380.919
shirt                  1481.1127   4564.977      0.324      0.747   -7713.22
2    1.07e+04
fireworks              2.063e+04   8331.793      2.477      0.017    3853.31
0    3.74e+04
bobblehead             9717.7744   3170.781      3.065      0.004    3331.49
3    1.61e+04
month_AUG              7073.4047   7644.848      0.925      0.360   -8324.11
0    2.25e+04
month_JUL              4567.7936   6208.828      0.736      0.466   -7937.42
8    1.71e+04
month_JUN              1712.1815   1.08e+04      0.158      0.875   -2.01e+0
4    2.35e+04
month_MAY              2743.6115   5943.158      0.462      0.647   -9226.52
3    1.47e+04
month_OCT              2433.8220   9039.116      0.269      0.789   -1.58e+0
4    2.06e+04
month_SEP              2301.9535   7626.007      0.302      0.764   -1.31e+0
4    1.77e+04
day_of_week_Monday     1.789e+04   9197.764      1.945      0.058    -638.63
6    3.64e+04
day_of_week_Saturday   2.217e+04   8717.029      2.543      0.014    4614.69
6    3.97e+04
day_of_week_Sunday     1.995e+04   9229.889      2.161      0.036    1358.22
2    3.85e+04
```

```
day_of_week_Thursday    1.89e+04   9196.691      2.055      0.046     374.57
9    3.74e+04
day_of_week_Tuesday    2.649e+04   9278.789      2.855      0.006    7799.67
4    4.52e+04
day_of_week_Wednesday  1.785e+04   8464.933      2.109      0.041     804.41
8    3.49e+04
opponent_Astros       -1.293e+04   1.15e+04     -1.129      0.265    -3.6e+0
4    1.01e+04
opponent_Braves       -1.221e+04   1.17e+04     -1.042      0.303   -3.58e+0
4    1.14e+04
opponent_Brewers      -1.365e+04   1.16e+04     -1.173      0.247   -3.71e+0
4    9791.866
opponent_Cardinals    -6344.3490   1.11e+04     -0.571      0.571   -2.87e+0
4     1.6e+04
opponent_Cubs         -6356.4325   1.17e+04     -0.543      0.590      -3e+0
4    1.72e+04
opponent_Giants       -1.026e+04    1.1e+04     -0.934      0.355   -3.24e+0
4    1.19e+04
opponent_Marlins      -1.192e+04   1.15e+04     -1.040      0.304    -3.5e+0
4    1.12e+04
opponent_Mets         -2467.2430   6048.281     -0.408      0.685   -1.46e+0
4    9714.621
opponent_Nationals       60.9403   1.21e+04      0.005      0.996   -2.42e+0
4    2.43e+04
opponent_Padres       -6681.5217   1.01e+04     -0.661      0.512    -2.7e+0
4    1.37e+04
opponent_Phillies     -8082.6966    1.1e+04     -0.737      0.465   -3.02e+0
4     1.4e+04
opponent_Pirates      -7683.9571   1.22e+04     -0.632      0.531   -3.22e+0
4    1.68e+04
opponent_Reds         -1.302e+04   1.13e+04     -1.156      0.254   -3.57e+0
4    9668.725
opponent_Rockies      -1.068e+04   1.09e+04     -0.984      0.331   -3.26e+0
4    1.12e+04
opponent_Snakes       -1.366e+04   1.06e+04     -1.285      0.205   -3.51e+0
4    7749.655
opponent_White Sox     -795.8206   5724.914     -0.139      0.890   -1.23e+0
4    1.07e+04
skies_Cloudy            271.1664   2350.539      0.115      0.909   -4463.06
3    5005.396
day_night_Night       -3052.1419   3544.265     -0.861      0.394   -1.02e+0
4    4086.375
=============================================================================
==
Omnibus:                       7.689   Durbin-Watson:                     2.4
40
Prob(Omnibus):                 0.021   Jarque-Bera (JB):                  7.1
35
Skew:                          0.683   Prob(JB):                         0.02
82
Kurtosis:                      3.497   Cond. No.                      4.54e+
03
=============================================================================
==

Notes:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
[2] The condition number is large, 4.54e+03. This might indicate that there
are
strong multicollinearity or other numerical problems.
```

To isolate effects while controlling for confounders, we use OLS regression with attendance as the dependent variable.

Predictors include temperature, binary promotions, and dummies for month, day, opponent, etc.

The model (R^2=0.709) explains 71% of variance, though adjusted R^2 (0.482) suggests some overfit due to many predictors (35) relative to observations (81).

Significant factors: bobbleheads (+9,718 fans, p=0.004), fireworks (+20,632, p=0.017—stronger than raw averages suggest), and certain days (e.g., Tuesday +26,490 vs. Friday baseline).

Opponents and months are mostly insignificant after controls, but rivals remain positive. This confirms promotions' causal potential, beyond correlations from earlier steps.

Recommendations to Management

1. The easiest win should be to expand Bobblehead Giveaways. The data shows a roughly +9,718 fan boost. Increase the number of giveaway days from 11 to 20 per season. Plan them on the currently lowest attendance days to boost those (Monday and Wednesday). Estimated revenue increase: + $485,900/game at$50/ticket average.
2. Try to do more fireworks on those lower attended days of the week.
3. Add more promotions in April/May and September/October the slowest times and promote midweek games aggressively.
4. Try new promotions and run testing on those.
   All of the together could increase average attendance to 45,000+, adding millions in revenue.