Analysis of Bobbleheads & Baseball Attendance

Louis Bailey

Contents

1 | Introduction

- In this notebook I will be using data on the Los Angeles Dodgers Major League Baseball team. The data will be explored with the goal of finding insights which can inform management on how to improve attendance.

- To keep this notebook from being overly long, this will be a partial analysis focused on only two variables related to attendance: bobbleheads and day of the week

- The dataset was already checked, and it appears no cleaning is necessary.

2 | Libraries & Data

In [1]:
```
import pandas as pd
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt

from scipy import stats
from scipy.stats import skew
from scipy.stats import shapiro
import statsmodels.stats.multicomp as multi
```

In [2]:
```
data = pd.read_csv('dodgers-2022.csv')
print(f'Shape:{data.shape}')
data.head()
```

```
Shape:(81, 12)
```

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

3 | Variable Encoding

> To determine correlation, the ordinal variables 'month' and 'day_of_week' need to be encoded

---

In [3]:
```
encodings = {
    "month": {'APR':4, 'MAY':5, 'JUN':6, 'JUL':7, 'AUG':8, 'SEP':9, 'OCT':10},
    "day_of_week": {'Monday':1,'Tuesday':2, 'Wednesday':3, 'Thursday':4, 'Friday':5, 'Saturday':6, 'Sunday':7},
}
```

In [4]:
```
pd.set_option("future.no_silent_downcasting", True) #silence future warning

data = data.replace(encodings)
```

> The nominal variables need be transformed into dummy vaiables

---

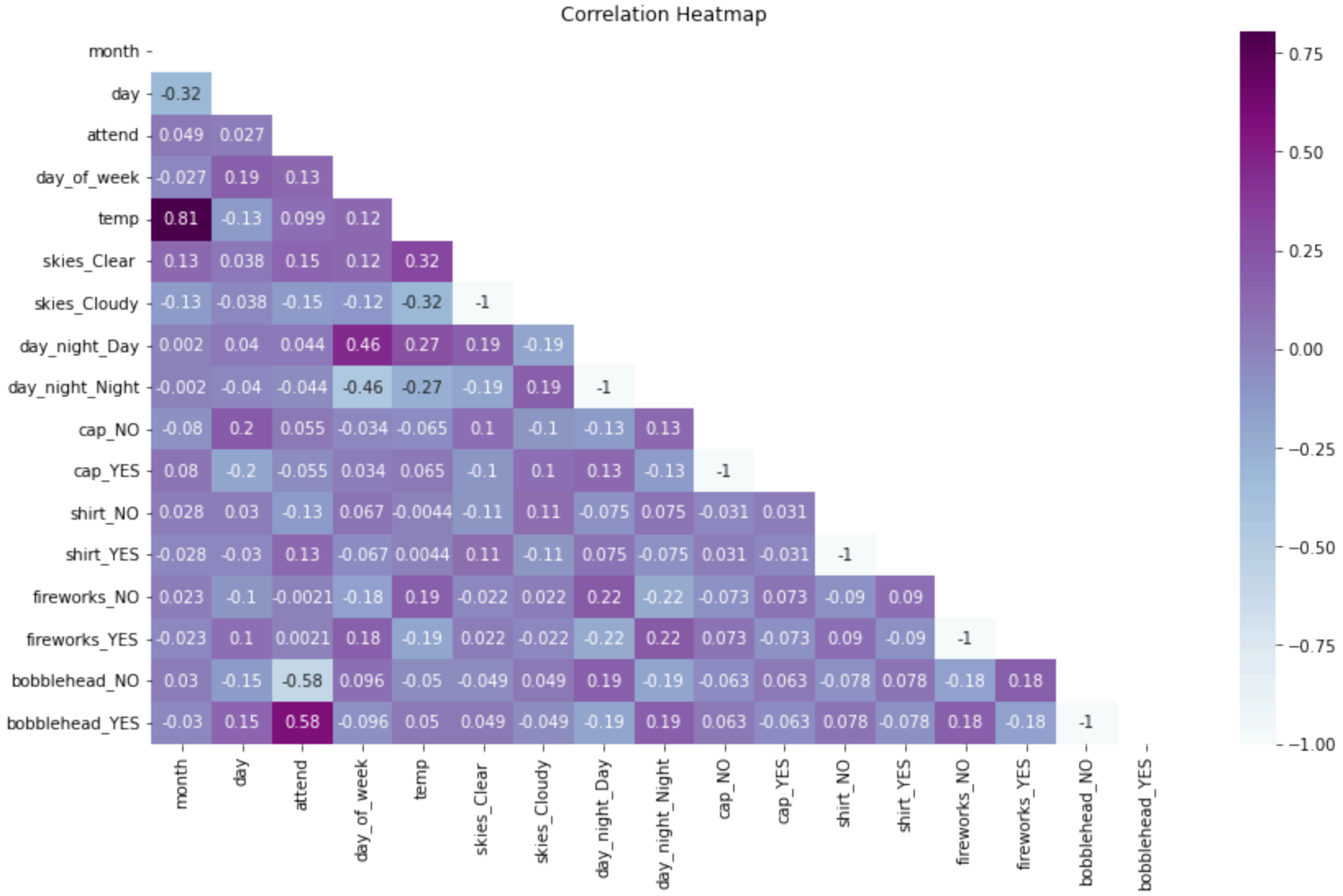
In [5]:
```
data = pd.get_dummies(data, columns=['skies','day_night','cap','shirt','fireworks','bobblehead'])
data.head()
```

Out[5]:

	month	day	attend	day_of_week	opponent	temp	skies_Clear	skies_Cloudy	day_night_Day	day_night_Night	cap_NO	cap_YES	shirt_NO	shirt_YES
0	4	10	56000	2	Pirates	67	True	False	True	False	True	False	True	False
1	4	11	29729	3	Pirates	58	False	True	False	True	True	False	True	False
2	4	12	28328	4	Pirates	57	False	True	False	True	True	False	True	False
3	4	13	31601	5	Padres	54	False	True	False	True	True	False	True	False
4	4	14	46549	6	Padres	57	False	True	False	True	True	False	True	False

4 | Correltions

In [6]:
```
#
corr_data = data.loc[:,~data.columns.str.startswith('opponent')]
plt.figure(figsize=(14,8))
mask = np.triu(np.ones_like(corr_data.corr()))
sns.heatmap(corr_data.corr(), annot=True,mask=mask, cmap='BuPu')
plt.title('Correlation Heatmap')
plt.show()
```



> From the correlation heatmap, there does not appear to be any redundant/collinear variables. Regarding attendance, bobblehead has the strongest correlation. Also of possible linear importance are shirts, skies, day of the week, and temp.

> Bobbleheads <

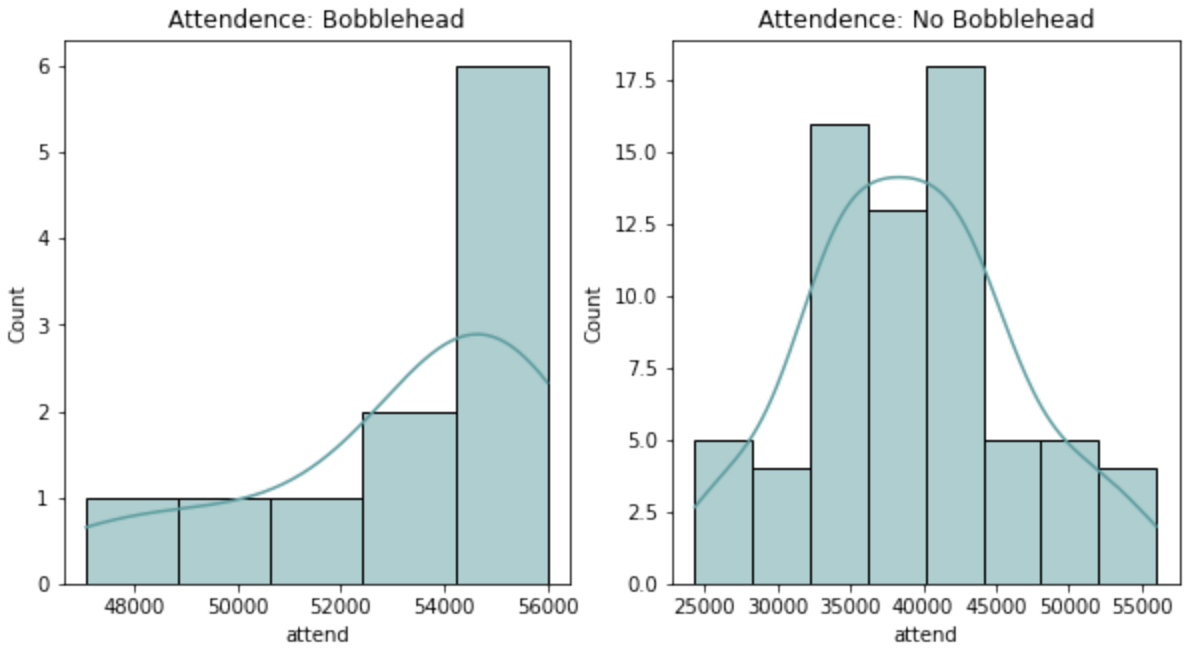
5 | Bobblehead Plots

In [7]:
```
#groups
bobblehead    = data[data.bobblehead_YES].attend
no_bobblehead = data[data.bobblehead_NO].attend
```

In [8]:
```
len(bobblehead)
```

Out[8]:
```
11
```

In [9]:
```
#
fig, (ax1, ax2) = plt.subplots(1,2, figsize=(10,5))

sns.histplot(bobblehead, ax=ax1, color='cadetblue', kde=True)
ax1.set_title('Attendence: Bobblehead')

sns.histplot(no_bobblehead, ax=ax2, color='cadetblue', kde=True)
ax2.set_title('Attendence: No Bobblehead')
plt.show()
```



> It appears there is a difference in attendance when bobbleheads are sold. To check whether this difference is significant, and how large that difference is, Welch's t-test and Hedge's g will be used. Welch's t-test is being used because the two groups do not have equal variance.
>
> Before that, these distributions need to be checked for normality.

---

6 | Bobblehead Normality

6.1 | Shapiro-Wilk Test

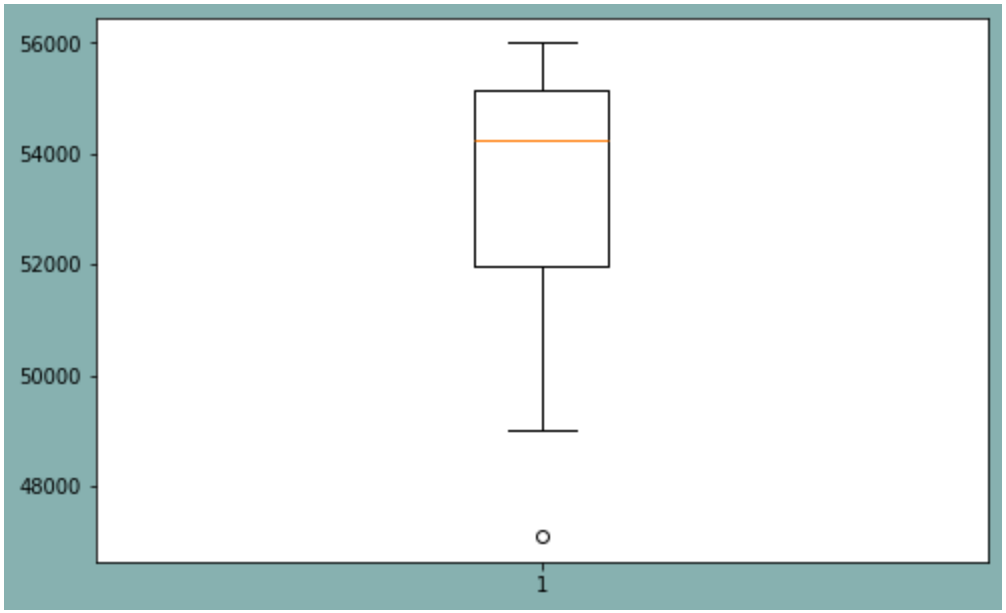
In [10]:
```
print(f'Bobblehead Shapiro p-value:\n{shapiro(bobblehead)[1]}')
print(f'\nNo Bobblehead Shapiro p-value:\n{shapiro(no_bobblehead)[1]}')
```

```
Bobblehead Shapiro p-value:
0.037406764924526215

No Bobblehead Shapiro p-value:
0.5448899865150452
```

> For the No_Bobblehead distribution, we fail to reject the null hypothesis, but for the Bobblehead distribution we reject the null hypothesis.
>
> The Bobblehead distibution will be checked for outliers, which if dropped may make the distribution normal.

6.2 | Checking for Outliers in Bobblehead

In [11]:
```
#

plt.figure(facecolor='#87B1B0', figsize=(8,5))
plt.boxplot(bobblehead)
plt.show()
```



In [12]:
```
#dropping outlier from bobblehead
threshold = 49000
bobblehead = bobblehead[bobblehead > threshold]
```

In [13]:
```
len(bobblehead)
```

Out[13]:
```
10
```

> Next these distributions will be tested again using the Shapiro Wilk test

---

6.3 | Shapiro-Wilk Test Again

In [14]:
```
print(f'Bobblehead Shapiro:\n{shapiro(bobblehead)[1]}')
print(f'\nNo Bobblehead Shapiro:\n{shapiro(no_bobblehead)[1]}')
```

```
Bobblehead Shapiro:
0.07290858030319214

No Bobblehead Shapiro:
0.5448899865150452
```

> For both the Bobblehead and No_Bobblehead distributions, since the p-value is greater than the chosen alpha of 0.05, we fail to reject the null hypothesis and assume the distributions are not significantly different from a normal distribution.

---

7 | Welch's t-test and Hedge's g

7.1 | Defining Functions

In [15]:
```
def t_test(group1, group2):
    t_stat, p_value = stats.ttest_ind(group1, group2, equal_var=False)
    print(f'The p-value is {p_value}')

def Hedges_g(group1, group2):
    diff = group1.mean() - group2.mean()
    var1 = group1.var()
    var2 = group2.var()
    avg_std = np.sqrt((var1 + var2) / 2)
```

```
    g = diff / avg_std
    return g
```

7.2 | Running Tests

In [16]:
```
t_test(bobblehead, no_bobblehead)

print(f"\nHedge's g: {Hedges_g(bobblehead, no_bobblehead)}")
```

```
The p-value is 4.513292312776158e-17

Hedge's g: 2.7530915168819803
```

7.3 | Results

t-test
p-value ~ 0

Hedge's g
~ 2.75

> Given that the p-value is less than the chosen alpha of 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference between the means of the two groups.
>
> A Hedge's g of 2.75 is very large. It indicates that the mean attendance at games where bobbleheads are sold is 2.75 standard deviations higher than the mean attendance at games where bobbleheads are not sold. That works out to an average of around 22,818 more attendees when bobbleheads are sold.
>
> Since the lowest value (the outlier) of the Bobblehead distribution was dropped to meet the reqirements of normality, this estimation may be a little high.

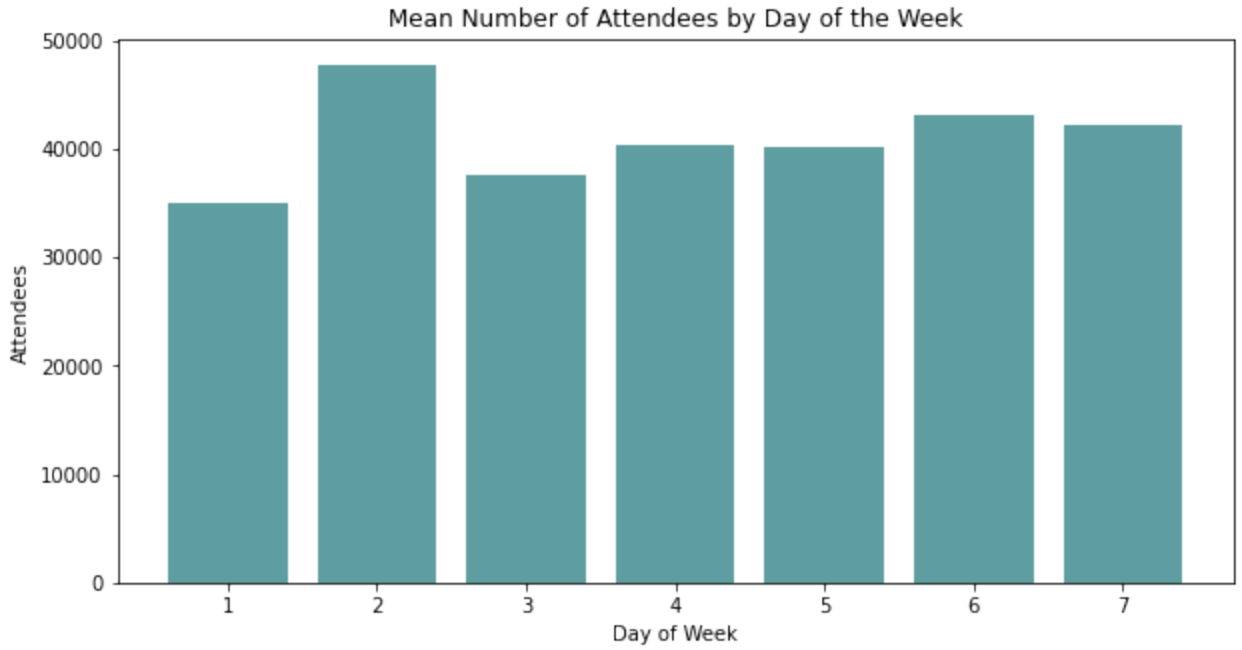
In [17]:
```
round(data.attend.std()*2.75,0)
```

Out[17]:
```
22818.0
```

>Day of the Week<

8 | Day of the Week Plot

In [18]:
```
#
plt.figure(figsize=(10,5))
plt.bar(x=['1','2','3','4','5','6','7'], height=data.groupby('day_of_week').attend.mean(), color='cadetblue')
plt.xlabel('Day of Week')
plt.ylabel('Attendees')
plt.title('Mean Number of Attendees by Day of the Week')
plt.show()
```



> It appears there are differences in mean attendence by day of the week. To check whether any of these differences are significant, Tukey's HSD test will be used. Before that, the Attendance by Day of the Week distributions need to be checked for homogeneity of variance and normality.

---

9 | Day of the Week Homogeneity of Variance, Normality & Tukey's HSD

9.1 | Levene's Test

In [19]:
```
monday, tuesday, wednesday, thursday, friday, saturday, sunday = [data[data.day_of_week==i].attend for i in range(1,8)]
```

In [20]:
```
stats.levene(monday, tuesday, wednesday, thursday, friday, saturday, sunday)
```

Out[20]:
```
LeveneResult(statistic=1.392112258774036, pvalue=0.22915168342167477)
```

> Since the p-value is greater than the chosen alpha of 0.05 we fail to reject the null hypothesis of equal variances. This suggests that the assumption of homogeneity of variance is met for these groups.

9.2 | Shapiro-Wilk Test

In [21]:
```
print(shapiro(monday))
print(shapiro(tuesday))
print(shapiro(wednesday))
print(shapiro(thursday))
print(shapiro(friday))
print(shapiro(saturday))
```

```
ShapiroResult(statistic=0.9197006225585938, pvalue=0.2834431529045105)
ShapiroResult(statistic=0.8718740940093994, pvalue=0.0554166808724403 4)
ShapiroResult(statistic=0.9552205204963684, pvalue=0.7140557169914246)
ShapiroResult(statistic=0.8962111473083496, pvalue=0.38930022716522217)
ShapiroResult(statistic=0.9818068146705627, pvalue=0.9871065616607666)
ShapiroResult(statistic=0.9463753700256348, pvalue=0.5445184707641602)
```

> For all of the distributions, the p-value is greater than the chosen alpha of 0.05, so we fail to reject the null hypothesis and assume the distributions are not significantly different from a normal distribution.

9.3 | Tukey's HSD test

In [22]:
```
f_value, p_value = stats.f_oneway(monday, tuesday, wednesday, thursday, friday, saturday, sunday)
print(f'f-value: {f_value}, p-value: {p_value}')

if p_value < 0.05:
    mc = multi.MultiComparison(data.attend, data.day_of_week)
    result = mc.tukeyhsd()

    print(result)
    print(mc.groupsunique)
```

```
f-value: 3.6440323261932344, p-value: 0.0031850342326589344
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================
group1 group2   meandiff  p-adj     lower      upper    reject
-------------------------------------------------------------
     1      2  12775.5641 0.0013   3578.4993 21972.6289   True
     1      3      2619.5  0.979  -6759.7025 11998.7025  False
     1      4   5441.7333 0.8265  -6787.2506 17670.7173  False
     1      5   5151.2564 0.6198  -4045.8084 14348.3212  False
     1      6   8107.2564 0.1202  -1089.8084 17304.3212  False
     1      7   7303.1795 0.2105  -1893.8853 16500.2442  False
     2      3 -10156.0641  0.021 -19353.1289  -958.9993   True
     2      4  -7333.8308 0.5269 -19423.6864  4756.0248  False
     2      5  -7624.3077 0.1522  -16635.554  1386.9386  False
     2      6  -4668.3077 0.7014  -13679.554  4342.9386  False
     2      7  -5472.3846 0.5256 -14483.6309  3538.8617  False
     3      4   2822.2333 0.9922  -9406.7506 15051.2173  False
     3      5   2531.7564 0.9805  -6665.3084 11728.8212  False
     3      6   5487.7564 0.5467  -3709.3084 14684.8212  False
     3      7   4683.6795 0.7178  -4513.3853 13880.7442  False
     4      5   -290.4769    1.0 -12380.3325 11799.3787  False
     4      6   2665.5231 0.9939  -9424.3325 14755.3787  False
     4      7   1861.4462 0.9992 -10228.4094 13951.3018  False
     5      6      2956.0 0.9537  -6055.2463 11967.2463  False
     5      7   2151.9231 0.9907  -6859.3232 11163.1694  False
     6      7   -804.0769    1.0  -9815.3232  8207.1694  False
-------------------------------------------------------------
[1 2 3 4 5 6 7]
```

9.4 | Resuls

From the adjusted p values, it appears there is a significant difference between Tuesday and Monday, and Tuesday and Wednesday.

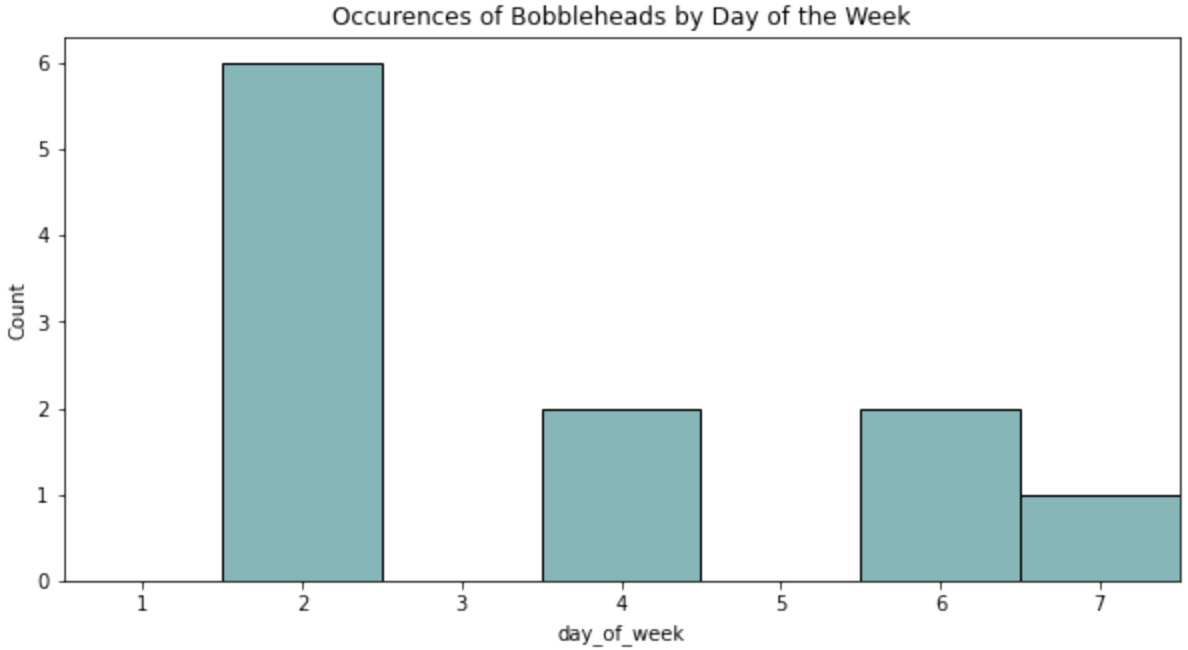
On average there are about 12,775 more attendees on Tuesday than on Monday.

On average there are about 10,156 more attendees on Tuesday than on Wednesday.

---

10 | Bobbleheads and Day of the Week Plot

In [25]:
```
#
bobblehead_day_of_week = data[data.bobblehead_YES==True].day_of_week

plt.figure(figsize=(10,5))
sns.histplot(bobblehead_day_of_week, color='cadetblue', discrete=True)
plt.xlim(0.5,7.5)
plt.title('Occurences of Bobbleheads by Day of the Week')
plt.show()
```



It appears that bobbleheads have mostly been sold on Tuesdays, and never on Mondays or Wednesdays. This could explain Tuesdays having significantly more attendees than Mondays and Wednesdays.

11 | Conclusion

- In the analysis above statistically significant differences in attendance were found between what day of the week a game falls on and whether bobbleheads are sold.

- From the results it appears that selling bobbleheads dramatically increases the number of attendees on average.

- Furthermore, it appears that attendance is on average significantly higher on Tuesdays than on Mondays or Wednesdays.

- Finally, it was observed that bobbleheads have mostly been sold on Tuesdays, and never on Mondays or Wednesdays which could explain the significant difference in attendance.

- My tentative recommendation to management is to try selling bobbleheads on lower attendance days Monday and Wednesday to increase attendance. I would need to explore this dataset more before settling on recommendations.

- For example, one could check whether the other promotional items: fireworks, shirts and caps have a significant impact on attendance. Additionally, attendance numbers could be analyzed in relation to 'month', 'day', 'day of the week', 'opponent', 'temp', and 'skies'.