Rodriguez_Felipe_DSC630_Assignment_3.2_Code

September 17, 2023

[1]:
```
# Import libraries
import pandas as pd
import matplotlib.pyplot as plt
```

[2]:
```
# Read in data
df = pd.read_csv("dodgers-2022.csv")
```

The analysis of this data will include understanding the relationship between attendance and the other variables within the data. Attendance is an important factor for the Los Angeles Dodgers and understanding the driving factors in attendance is crucial to forecast profits for the upcoming seasons. By analyzing the variables that influence attendance, a recommendation can be made to further improve areas that will likely affect the attendance for a baseball game.

[3]:
```
# Show data
df.head()
```

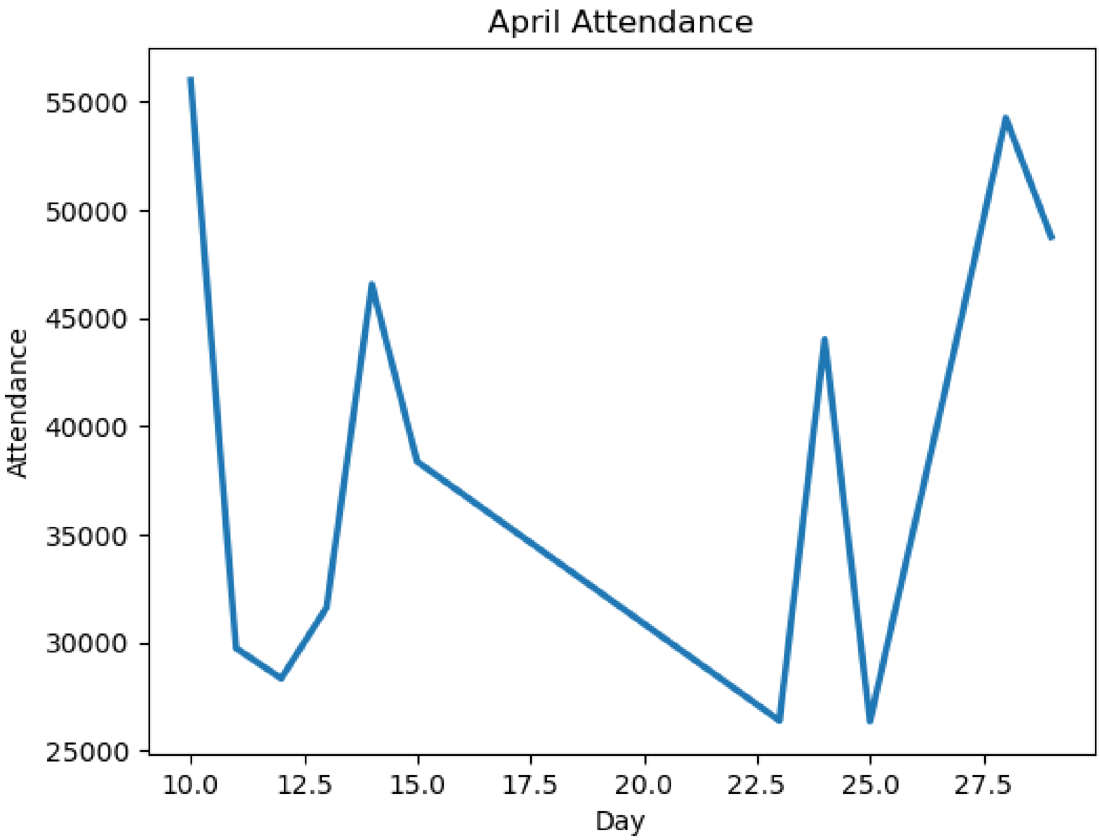
[3]:
```
  month  day  attend day_of_week opponent  temp   skies day_night cap shirt  \
0   APR   10   56000     Tuesday  Pirates    67  Clear         Day  NO    NO
1   APR   11   29729   Wednesday  Pirates    58  Cloudy      Night  NO    NO
2   APR   12   28328    Thursday  Pirates    57  Cloudy   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

  fireworks bobblehead
0        NO         NO
1        NO         NO
2        NO         NO
3       YES         NO
4        NO         NO
```

[4]:
```
# Set X and Y for April Attendance
x = df[df['month'] == "APR"]['day']
y = df[df['month'] == "APR"]['attend']
```

To gain some insight on the data, the attendance by day for the months of April and May are graphed to show how the attendance changes over time.

[5]:

```python
# Setup plot
fig, ax = plt.subplots()
# Plot X and Y
ax.plot(x, y, linewidth=2.5)
# Set labels for graph
plt.title('April Attendance')
plt.xlabel('Day')
plt.ylabel('Attendance')
# Display Graph
plt.show()
```



[6]:

```python
# Calculate total attendance for April
april_total = df[df['month'] == "APR"]['attend'].sum()
print('Total attendance for April', april_total)
```

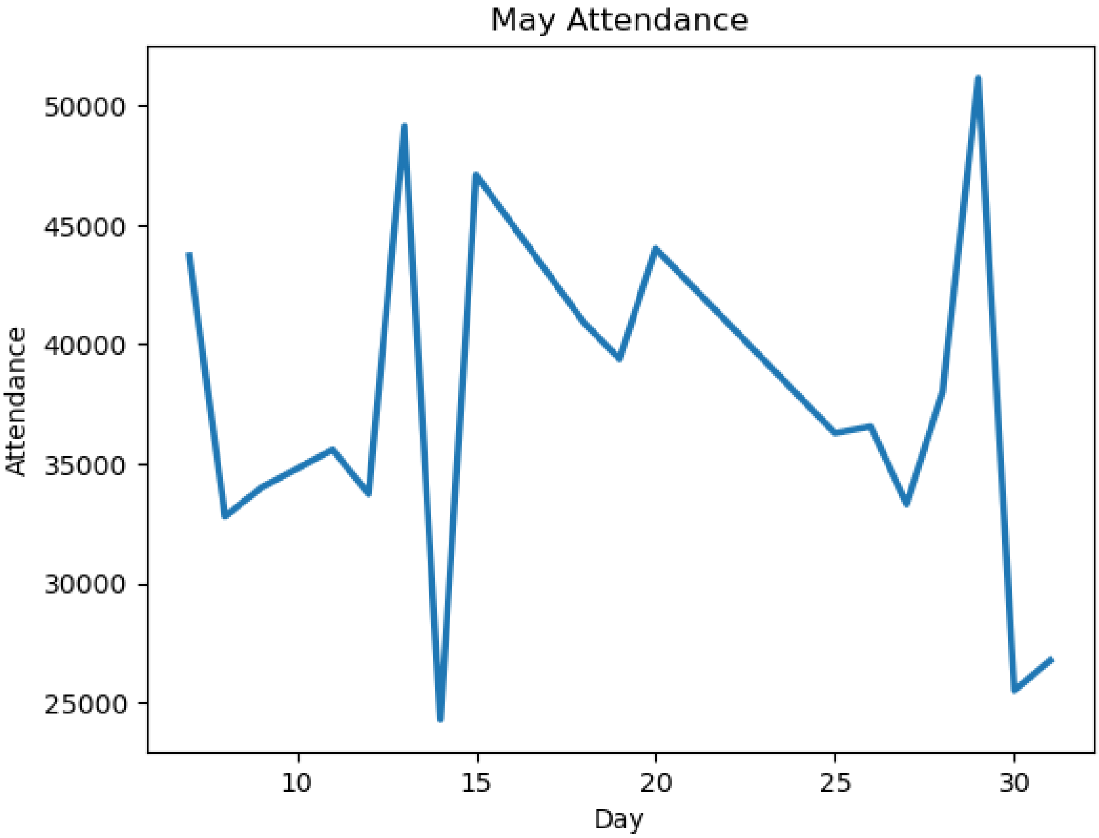
```
Total attendance for April 475103
```

[7]:

```python
# Set X and Y for May Attendance
x = df[df['month'] == "MAY"]['day']
y = df[df['month'] == "MAY"]['attend']
```

[8]:

```
# Setup plot
fig, ax = plt.subplots()
# Plot X abd Y
ax.plot(x, y, linewidth=2.5)
# Set labels for graph
plt.title('May Attendance')
plt.xlabel('Day')
plt.ylabel('Attendance')
# Display Graph
plt.show()
```



[9]:

```
# Calculate total attendance for May
may_total = df[df['month'] == "MAY"]['attend'].sum()
print('Total attendance for May', may_total)
```

```
Total attendance for May 672223
```

When comparing attendance for both months, there is no similarities between the two. It can be noted that the May attendance is higher than April attendance. This is confirmed by adding attendance for each month. For April, total attendance was 475,103 and for May it was 672,223. One reason May is higher in attendance is that the data contains a full month for May. The data starts from April 10th which omits part of that month.

[10]:

```
from sklearn.preprocessing import LabelEncoder
```

[11]:

```
# Get the list of categorical columns
categorical_columns = df.select_dtypes(include=['object']).columns

# Apply label encoding on each categorical column
for column in categorical_columns:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column])
```

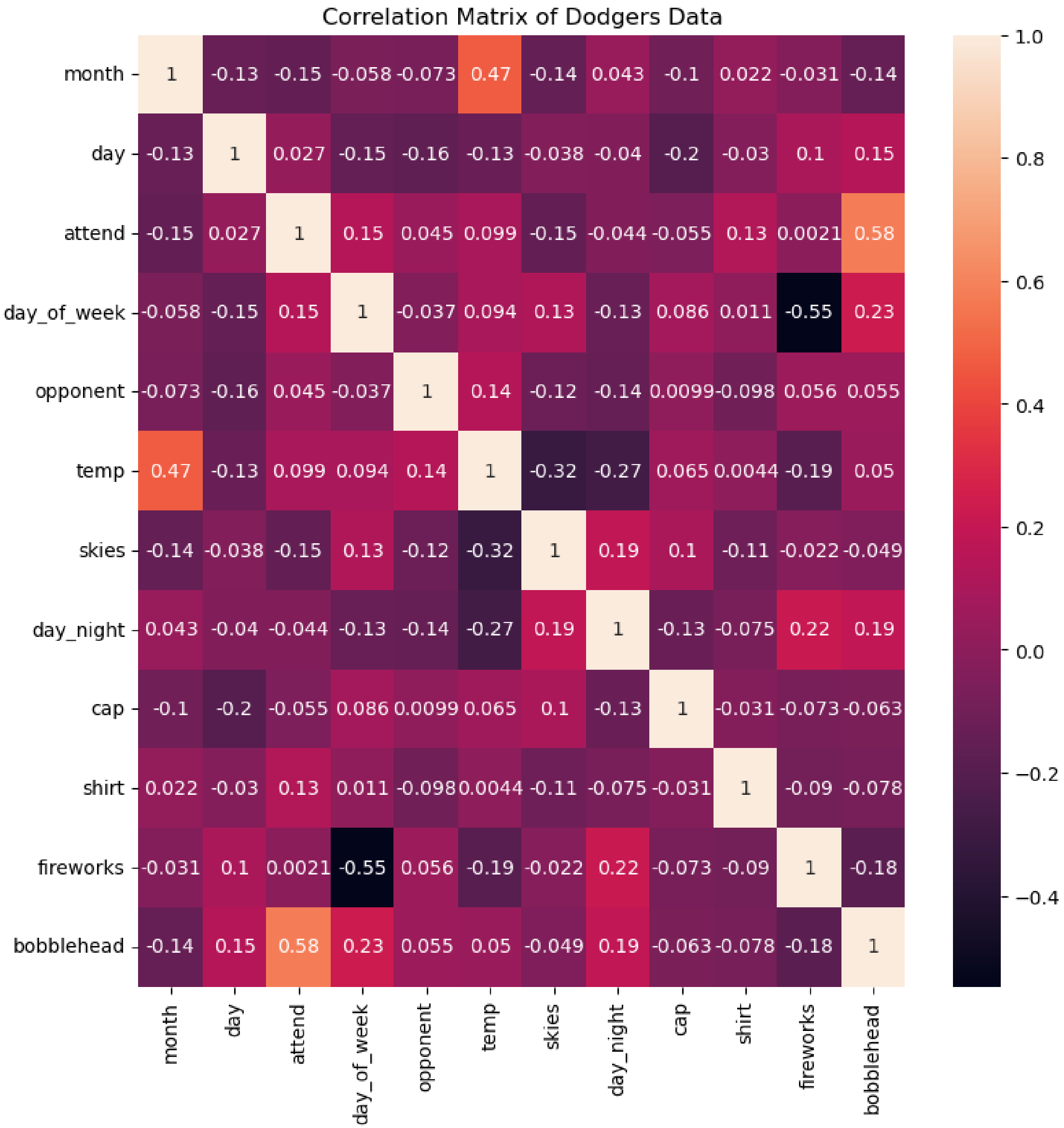
[12]:

```
# Import libraries
import seaborn as sn
```

A great way to view correlation is by creating a correlation matrix, this will display the values that influence each other, and we can focus on attendance to see which variable affects it the most.

[15]:

```
# Create correlation matrix of data
fig, ax = plt.subplots(figsize=(9,9))
corr_matrix = df.corr()
sn.heatmap(corr_matrix, annot=True)
plt.title('Correlation Matrix of Dodgers Data')
# Display Matrix
plt.show()
```



The correlation matrix shows us that the variable that positively influences attendance the most is if there was a bobblehead during the game. The variable that negatively affects attendance is skies, however it is not possible to control this variable.

The next factor explored, will be understanding how days of the week play a role in the games. Although day of the week had a slight positive correlation on attendance, further exploration can uncover more details of this variable.

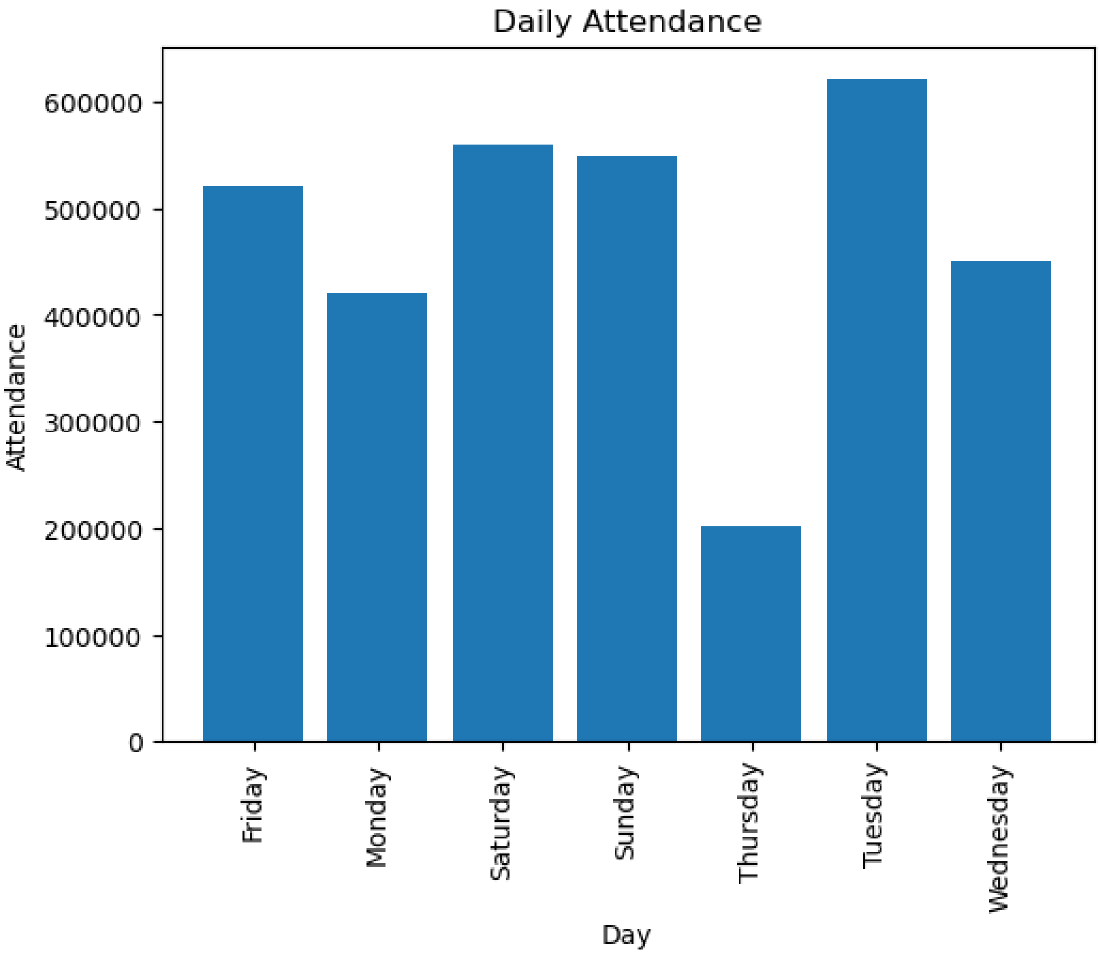
[16]:
```
# Read data
df = pd.read_csv("dodgers-2022.csv")
```

```
[17]: # Group data by day of week
      df_days = df.groupby('day_of_week')['attend'].sum().reset_index()
      df_days
```

```
[17]:   day_of_week  attend
      0      Friday  521520
      1      Monday  419588
      2    Saturday  559948
      3      Sunday  549495
      4    Thursday  202037
      5     Tuesday  620636
      6   Wednesday  451022
```

```
[18]: # Set X and Y for Daily Attendance
      x = df_days['day_of_week']
      y = df_days['attend']
```

```
[19]: # Setup plot
      fig, ax = plt.subplots()
      # Plot X and Y
      ax.bar(x, y, linewidth=2.5)
      # Set labels for graph
      plt.title('Daily Attendance')
      plt.xlabel('Day')
      plt.xticks(rotation=90)
      plt.ylabel('Attendance')
      # Display graph
      plt.show()
```



```
[20]: # Count how many games per day
      value_counts = df['day_of_week'].value_counts()
      value_counts
```

```
[20]: day_of_week
      Tuesday      13
      Friday       13
      Saturday     13
      Sunday       13
      Wednesday    12
      Monday       12
      Thursday      5
      Name: count, dtype: int64
```

When looking at attendance by day, the weekend days are noticeably the highest, but to our surprise, Tuesday has the most attendance of the games. When looking at the number of games for each day, Friday through Monday had around the same number of days, while Thursday had only 5 games, this explains why Thursdays had least attendance.

After analyzing the data, we have uncovered three recommendations for attendance. The first recommendation is to include the bobblehead in more games. This was uncovered in the correlation matrix, and it showed to have a positive correlation between attendance and the appearance of the bobblehead. The second recommendation is to redistribute the more games to Thursdays. Since this day already has the fewest number of games, spreading the games out to include more Thursdays can create an additional opportunity for fans to attend games. The last recommendation is to continue to promote Tuesday games since they drive the most attendance.