DSC630_KristaKnuckey_Week3

April 5, 2025

[ ]:
```
#Krista Knuckey
#Week3
```

[7]:
```
#importing libraries
```

[26]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

[9]:
```
dodgers_data = pd.read_csv('dodgers.csv')
dodgers_data.head()
```

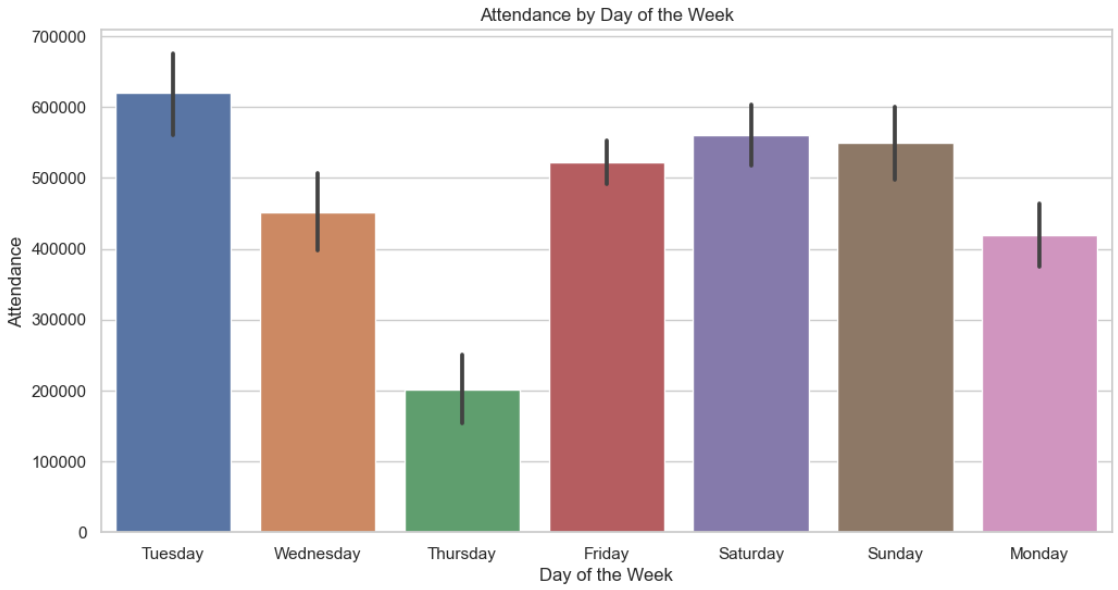
[9]:

	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

'm/M' Visualization 1- Attendance of game by Day of the Week to see what days are most popular

[11]:
```
plt.figure(figsize=(12,6))
sns.barplot(data=dodgers_data, x='day_of_week', y='attend', estimator=sum)
plt.title('Attendance by Day of the Week')
plt.xlabel('Day of the Week')
plt.ylabel('Attendance')
```

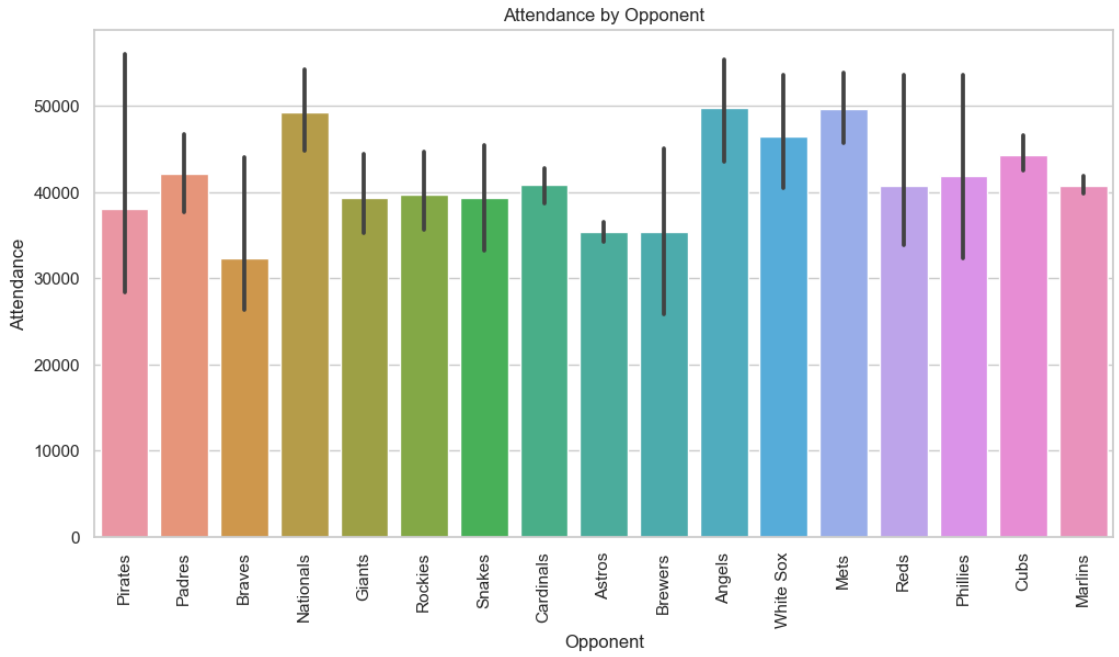
```
plt.show()
```



'm/M' We can see that Tuesday has the highest day of attendance, even when compared to weekend days. We do not have the data to support this, but I am curious to see if the price of the ticket on weekday vs a weekend is the reason for high attendance on a Tuesday.

'm/M' Visualization 2- Attendance by Opponent. Here we should see if there are any opponents that fans would like to see more than others, or if this affects high attendance on a weekday.

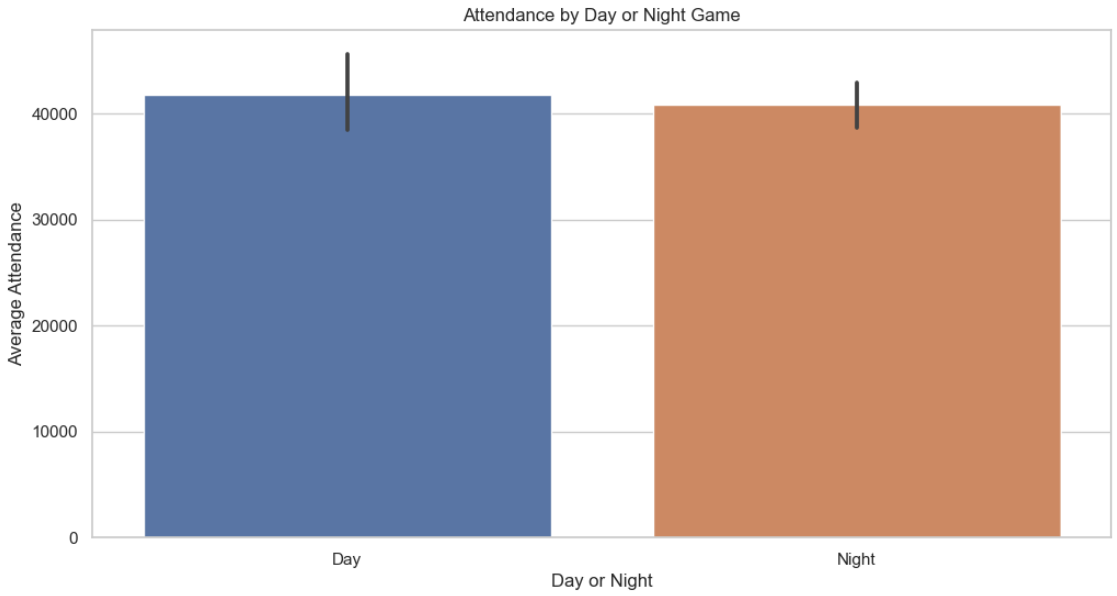
[18]:
```
plt.figure(figsize=(12,6))
sns.barplot(data=dodgers_data, x='opponent', y='attend')
plt.title('Attendance by Opponent')
plt.xlabel('Opponent')
plt.ylabel('Attendance')
plt.xticks(rotation=90)
plt.show()
```



'm/M' Through this visualization, we can see that the more popular games were against the Nationals, Angels, and Mets. However, there is relatively high attendance in all games against the opponents.

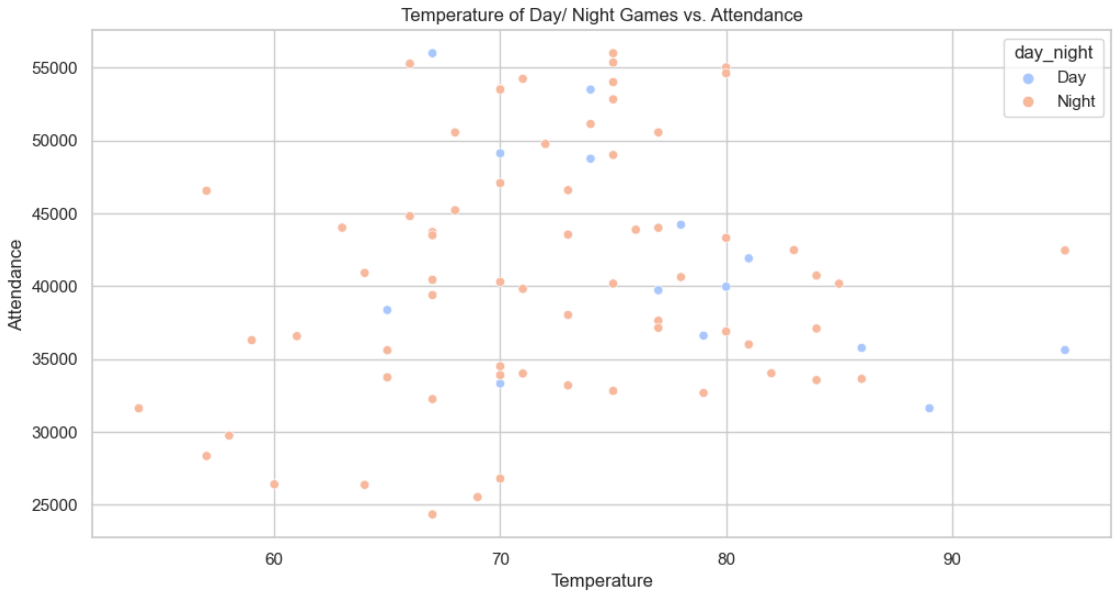
'm/M' Visualization 3- Weather and time of the game will have an impact on attendance. First, we are going to visualize attendance by the day/night game. Then compare that to the temperature of the day/ night games to see how weather has impacted as well.

[19]:
```
plt.figure(figsize=(12,6))
sns.barplot(data=dodgers_data, x='day_night', y='attend')
plt.title('Attendance by Day or Night Game')
plt.xlabel('Day or Night')
plt.ylabel('Average Attendance')
plt.show()
```



'm/M' From this visualization, we can see that whether the game is scheduled during the day or night does not have a huge impact on attendance as they are almost exactly equal. Next, we will see how the temperature affects this as night games can be "cold" vs day games can be "hot".

[20]:
```
plt.figure(figsize=(12,6))
sns.scatterplot(data=dodgers_data, x= 'temp', y= 'attend', hue= 'day_night',␣
↪palette= 'coolwarm')
plt.title('Temperature of Day/ Night Games vs. Attendance')
plt.xlabel('Temperature')
plt.ylabel('Attendance')
plt.show()
```

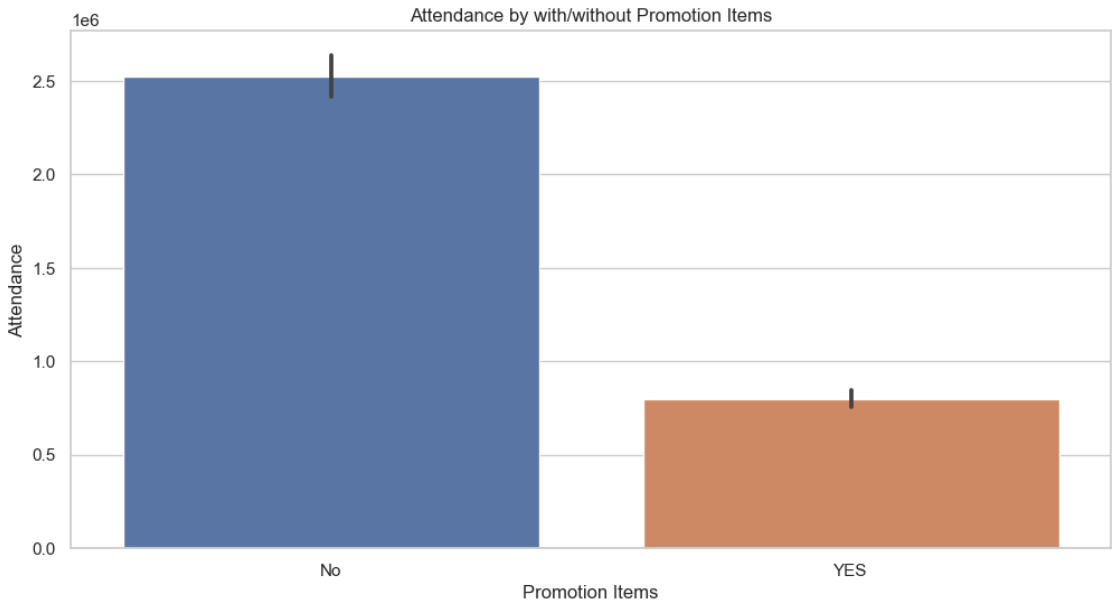


'm/M' From the visualization, we can see that there are more night games vs. day games. However, it appears that temperature does not impact attendance as much. We can see that attendance is at it's peak during more comfortable temperatures, like 70-80 degrees, but the outliers of the high 90's show that fans will still show up to games whether it's day or night.

'm/M' Visualization 4- Comparing attendance on games that gave out promotional items, like caps, bobbleheads, and shirts vs the games that did not.

[24]:
```
dodgers_data['promotion_item'] = dodgers_data[['cap', 'shirt', 'bobblehead']].␣
↪apply(lambda x:'YES' if 'YES' in x.values else 'No', axis=1)

plt.figure(figsize=(12,6))
sns.barplot(data=dodgers_data, x='promotion_item', y='attend', estimator=sum)
plt.title('Attendance by with/without Promotion Items')
plt.xlabel('Promotion Items')
plt.ylabel('Attendance')
plt.show()
```



'm/M' Creating a linear regression model based on the previous variables that visualizations were created for.

[33]:
```
features = ['day_of_week', 'opponent', 'temp', 'day_night', 'cap', 'shirt',␣
 ↪'bobblehead']
X = dodgers_data[features]
y = dodgers_data['attend']

categorical_features = ['day_of_week', 'opponent', 'day_night', 'cap', 'shirt',␣
 ↪'bobblehead']
encoder = OneHotEncoder(drop='first', sparse=False)
X_encoded = encoder.fit_transform(X[categorical_features])
encoded_feature_names = encoder.get_feature_names_out(categorical_features)

X_encoded_df = pd.DataFrame(X_encoded, columns=encoded_feature_names)
X_encoded_df['temp'] = X['temp'].values

X_train, X_test, y_train, y_test = train_test_split(X_encoded_df, y,␣
 ↪test_size=0.2, random_state=42)

lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
y_pred = lin_reg.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
```

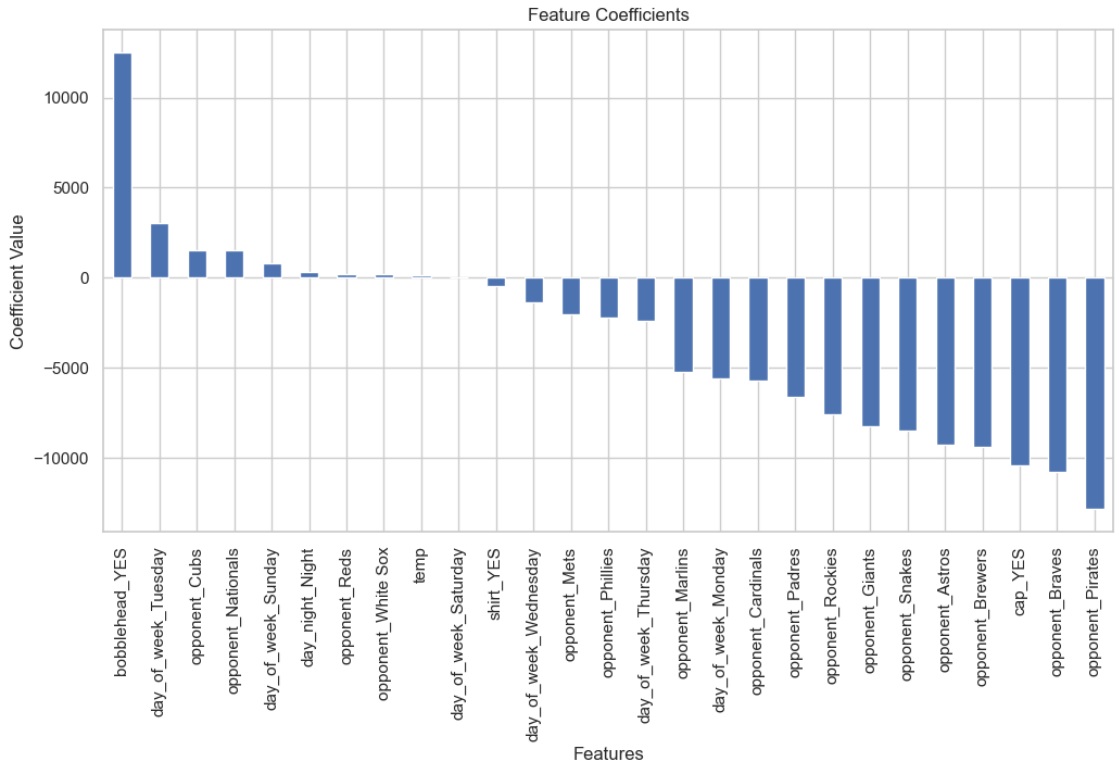
```
print(f'R-squared: {r2}')

coefficients = pd.Series(lin_reg.coef_, index=X_encoded_df.columns)
coefficients_sorted = coefficients.sort_values(ascending=False)

plt.figure(figsize=(12, 6))
coefficients_sorted.plot(kind='bar')
plt.title('Feature Coefficients')
plt.xlabel('Features')
plt.ylabel('Coefficient Value')
plt.show()
```

```
Mean Squared Error: 77005352.0526987
R-squared: 0.21680585107856576
```

```
/opt/anaconda3/lib/python3.11/site-
packages/sklearn/preprocessing/_encoders.py:868: FutureWarning: `sparse` was
renamed to `sparse_output` in version 1.2 and will be removed in 1.4.
`sparse_output` is ignored unless you leave `sparse` to its default value.
  warnings.warn(
```



'm/M' Conclusion: Although the visualizations of the data points gave great insights of attendance, the model did have a MSE and R-squared score that indicate the model does not fit as well with the points I was comparing. However, since there is a lot of variability in sports, teams, weather,

etc it is good to mention that this variability can also indicate whether a model can or will be successful.

After plotting the coefficients to see what would have a positive impact on higher attendance rates, we see that bobbleheads had a great impact on attendance. I would suggest to the team that they invest in bobblehead promotions and make special editions to entice fans more. Also, we see that Tuesdays and Sundays have higher attendance- I would suggest scheduling these promotion items on those days, as well as, scheduling opponents that fans are more interested in watching. From this visualization, we can see that the Cubs, Nationals, Red Sox, and White Sox have a positive impact on attendance. A surprising outcome of these results is that the weather did not have as great an impact as I would have thought. I am not an avid sports fan, so bad weather would affect whether or not I would want to attend. This is a great insight that this does not impact fans- they will attend to support the team they love!

[ ]: