PROJECT CHECK-IN

Jacob Kronlage, Xavier Hummer, and Ryan Perry

INITIAL ANALYSIS

- We developed preliminary hypotheses based on our research questions before analyzing the data:
  - We believed that **On-Base Percentage (OBP)** and **Slugging Percentage (SLG)** would be key performance drivers.
  - We expected that home runs would strongly indicate team **success in making the playoffs.**
  - We anticipated that teams with greater success would attract **higher home game attendance.**

SCRAPED ATTENDANCE

- We aimed to compare team **home attendance** with overall team success.
- To do this, we scraped **home attendance data** from an additional website (The Baseball Cube).
- After scraping, we merged the **home attendance** data into our primary dataset for analysis.

Total Home Attendance
2177617
2420171
2102240
3043003
2882756
...
762169
1071614
1590136
938516
729741

```
# Loop through each year from 1962 to 2012
for year in range(1962, 2013):
    print("Scraping:", year)
    url = f"https://www.thebaseballcube.com/content/mlb_att_year/{year}"
    driver.get(url)
    time.sleep(3)  # Wait for the page to load

    # Find the table and rows
    table = driver.find_element(By.TAG_NAME, "table")
    rows = table.find_elements(By.TAG_NAME, "tr")

    # Loop through the rows, skip the header
    for row in rows[1:]:
        cells = row.find_elements(By.TAG_NAME, "td")

        if len(cells) >= 5: # Check cells have values
            team = cells[0].text
            att = cells[2].text.replace(',', '')

            if team != "" and "League Total" not in team and att.isdigit()
                team_list.append(team)
                year_list.append(year)
                attendance_list.append(int(att))
```

MACHINE LEARNING

- We built a linear regression model to find which factors drive team wins.
- OBP shows the strongest positive relationship with winning.
- BA and SLG have much weaker impacts compared to OBP.
- Total Home Attendance has almost no effect on a team's success.

```
# Assign features and target variable

X = checkin_df[['OBP', 'BA', 'SLG', 'Total Home Attendance']]
y = checkin_df['Wins']

# Train/test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Assign and fit model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict
y_pred = model.predict(X_test)

# Show the score and MSE of the model
print('Mean Squared Error:', mean_squared_error(y_test, y_pred))
print('R-squared:', r2_score(y_test, y_pred))

# Display model coefficients
coefficients = pd.DataFrame({'Feature': X.columns, 'Coefficient': model.coef_})
print(coefficients)
```

```
Mean Squared Error: 94.23057447906525
R-squared: 0.23759058560753865
                 Feature  Coefficient
0                    OBP   278.288398
1                     BA     7.822252
2                    SLG    -5.875437
3  Total Home Attendance     0.000004
```

HYPOTHESIS TEST

- We used a two-sample t-test to determine if home runs differ significantly between playoff and non-playoff teams.
- The t-test was statistically significant (**T = 9.82**, **P < 0.0001**), indicating a real difference between the two groups.
- Playoff teams, on average, hit significantly more home runs than teams that missed the postseason.
- We used a t-test because we were comparing the means of two independent groups (playoff vs. non-playoff teams).

HYPOTHESIS TEST

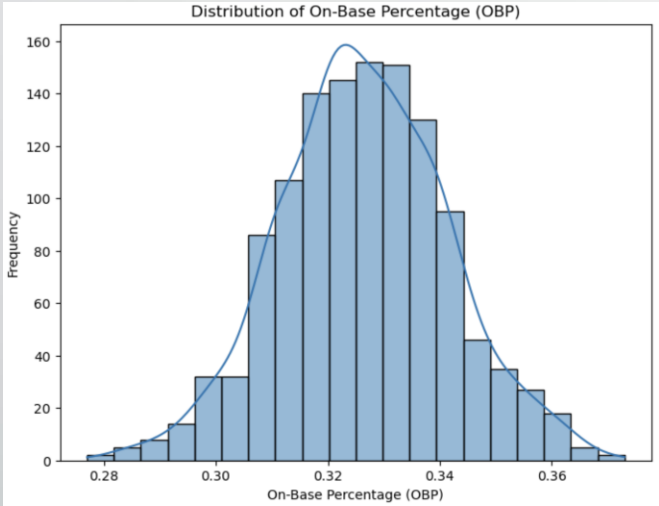
```python
# Hypothesis Test

# Change data type of column to numeric to perfrom tests
checkin_df['Home Runs'] = pd.to_numeric(checkin_df['Home Runs'], errors='coerce')

# The goal is to see if homeruns are significantly different for playoff vs non-playoff teams?

playoff_hr = checkin_df[checkin_df['Playoffs'] == 'Yes']['Home Runs'].dropna()
non_playoff_hr = checkin_df[checkin_df['Playoffs'] == 'No']['Home Runs'].dropna()

t_stat, p_value = stats.ttest_ind(playoff_hr, non_playoff_hr, equal_var=False)

print(f"T-statistic: {t_stat}")
print(f"P-value: {p_value}")

if p_value < 0.05:
    print("Result: Statistically significant difference in Home Runs between playoff and non-playoff teams.")
else:
    print("Result: No statistically significant difference in Home Runs between playoff and non-playoff teams.")
```

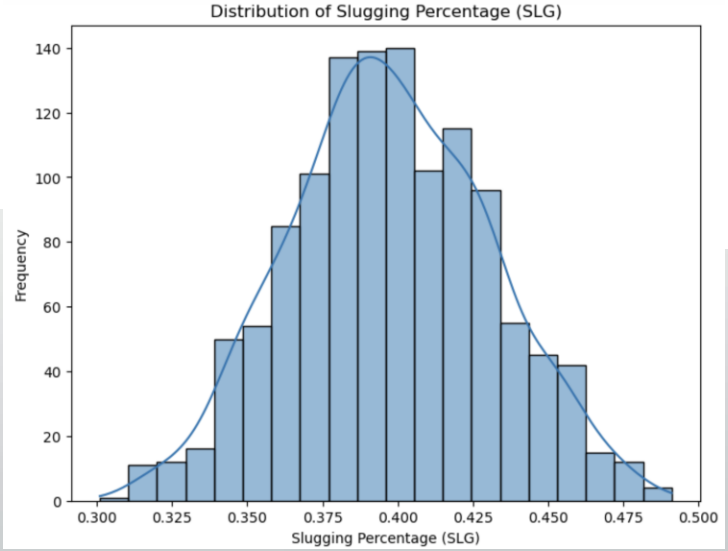
```
T-statistic: 9.823165656488838
P-value: 2.083950817803217e-20
Result: Statistically significant difference in Home Runs between playoff and non-playoff teams.
```

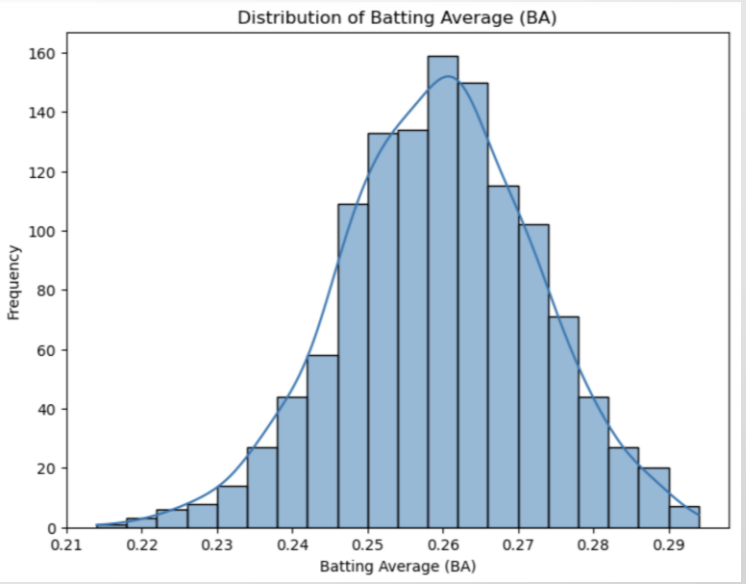
UNIVARIATE ANALYSIS

- We performed a univariate analysis to better understand the distribution of MLB team On-Base Percentage (OBP).
  - The median OBP was approximately **0.326**, with a mean of **0.326** and a standard deviation of **0.015**.
- We also analyzed Batting Average (BA) and Slugging Percentage (SLG).
  - The median BA was **0.260** (mean **0.259**, standard deviation **0.013**).
  - The median SLG was **0.396** (mean **0.397**, standard deviation **0.033**).
- These metrics provide a foundation to compare OBP, BA, and SLG when evaluating their importance to team success.

UNIVARIATE ANALYSIS







LOOKING AHEAD

- We need to continue answering our research questions by testing different modeling techniques beyond linear regression.
- We will run additional hypothesis tests and explore which features are most valuable for improving model performance.
- Our initial analysis showed we are on the right track, with a statistically significant t-test result.
- However, our linear regression model had a low R-squared value, suggesting we may need to add new features or remove weaker ones.
- We plan to conduct more bivariate analysis, using scatter plots and box plots, to better understand relationships between variables.
- We also need to further explore the role of Total Home Attendance, as it showed little predictive power for team wins in our linear regression model.

UPDATED DICTIONARY

- With the addition of Total Home Attendance, our dataset now includes 15 features.
- We feel confident that the features we have scraped are sufficient to proceed with our analysis.

Field	Type	Description
Team	Text	Abbreviation of full team name.
Team Name	Text	Full name of the team.
League	Text	Which league the team is in (American or National).
Year	Numeric	Which year the season was played in.
Games	Numeric	Number of games played.
Wins	Numeric	Number of wins.
Home Runs	Numeric	Total number of home runs hit collectively as a team.
RS	Numeric	Total number of runs scored as a team.
RA	Numeric	Total number of runs allowed as a team.
OBP	Numeric	*On Base Percentage* – The percentage of a team's plate appearances that result in a player reaching base.
SLG	Numeric	*Slugging Percentage* – A team's average number of total bases earned per at-bat.
BA	Numeric	*Batting Average* – The team's percentage of at-bats that result in a hit.
Playoffs	Logical	Yes or no if the team made the playoffs.
PlayoffsFinish	Numeric	Final team ranking in postseason – e.g., World Series Winner = 1.0
Attendance	Numeric	Total Number of Fans in Attendance to the game.