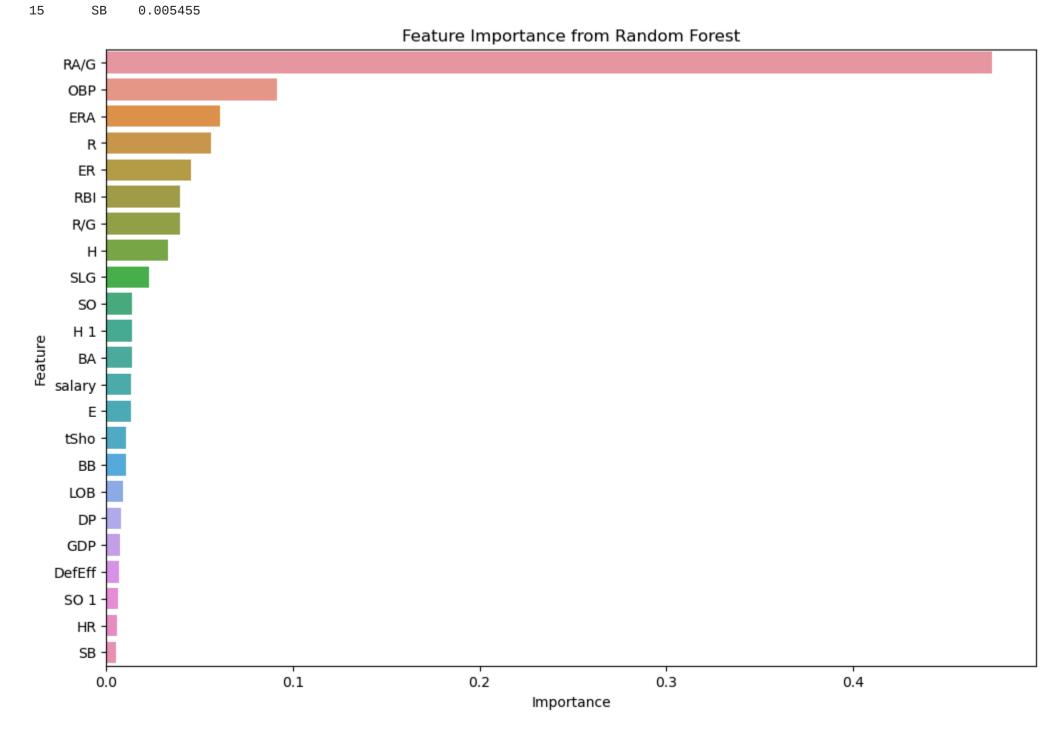
```
In [3]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.feature_selection import SelectKBest, f_regression
# Load the dataset
df = pd.read_csv('mlb_teams.csv')
# Display the first few rows of the dataset and column names
print(df.head())
print(df.columns)
# Check for missing values
print(df.isnull().sum())
# Inspect the column names to identify non-numeric columns
print(df.columns)
# Ensure column names are correctly identified
columns_to_drop = ['TeamName']
for col in df.columns:
     if 'season' in col.lower():
         columns_to_drop.append(col)
print(f"Columns to drop: {columns_to_drop}")
# Drop non-numeric columns if they exist (adjust as necessary based on actual column names)
df = df.drop(columns=columns_to_drop)
# If there are missing values, handle them (e.g., fill with mean, median, or drop)
df = df.fillna(df.mean())
# Verify all columns are numeric now
print(df.dtypes)
\# Define the features (X) and the target variable (y)
X = df.drop(columns=['W-L\%', 'W', 'L', 'WAR']) # Adjust based on your dataset's column names
y = df['W-L\%']
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Normalize/scale the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
# Feature selection using correlation
correlation_matrix = df.corr()
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
# Feature selection using SelectKBest
selector = SelectKBest(score_func=f_regression, k='all')
selector.fit(X_train_scaled, y_train)
feature_scores = pd.DataFrame({'Feature': X.columns, 'Score': selector.scores_})
feature_scores = feature_scores.sort_values(by='Score', ascending=False)
print(feature_scores)
# Fit a linear regression model
lr = LinearRegression()
lr.fit(X_train_scaled, y_train)
y_pred_lr = lr.predict(X_test_scaled)
 # Evaluate the linear regression model
print('Linear Regression RMSE:', np.sqrt(mean_squared_error(y_test, y_pred_lr)))
print('Linear Regression R^2:', r2_score(y_test, y_pred_lr))
# Fit a random forest regressor model
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X_train_scaled, y_train)
y_pred_rf = rf.predict(X_test_scaled)
# Evaluate the random forest model
print('Random Forest RMSE:', np.sqrt(mean_squared_error(y_test, y_pred_rf)))
print('Random Forest R^2:', r2_score(y_test, y_pred_rf))
# Feature importance from random forest
feature_importances = pd.DataFrame({'Feature': X.columns, 'Importance': rf.feature_importances_})
feature_importances = feature_importances.sort_values(by='Importance', ascending=False)
print(feature_importances)
# Plot feature importance
plt.figure(figsize=(12, 8))
sns.barplot(x='Importance', y='Feature', data=feature_importances)
plt.title('Feature Importance from Random Forest')
plt.show()
    TeamName RA/G DefEff
                               Ε
                                   DP
                                         W
                                              L W-L%
                                                          ERA tSho ...
                                                                           RBI \
   2018 ARI
              3.98
                     0.698
                              75
                                  152
                                        82
                                              80
                                                 0.506
                                                          3.72
                                                                   9
                                                                           658
                                                                      . . .
   2018 ATL
              4.06
                      0.709
                              80
                                  134
                                         90
                                              72
                                                  0.556
                                                          3.75
                                                                  11
                                                                     . . .
                                                                           717
   2018 BAL
              5.51
                     0.674
                             104
                                  159
                                        47
                                            115
                                                  0.290
                                                                  7
                                                                            593
                                                                      . . .
   2018 BOS
             3.99
                     0.693
                             77
                                  106
                                       108
                                              54 0.667
                                                          3.75
                                                                           829
                                                                  14 ...
                                                                  18 ...
   2018 CHC
             3.96
                     0.700
                            104
                                 155
                                        95
                                              68
                                                 0.583
                                                        3.65
                                                                           722
     SB S0 1
                  BA
                         0BP
                                SLG GDP
                                           LOB
                                                    salary
                                                             WAR
0
    79 1460
              0.235 0.310
                              0.397
                                     110
                                          1086
                                                143324597
                                                             34.1
1
    90 1290 0.257 0.324
                              0.417
                                      99 1143
                                                 130649395
                                                             40.8
    81 1412 0.239 0.298
                                     132 1027
                              0.391
                                                 127633703
                                                             11.4
3
   125 1253 0.268 0.339
                              0.453
                                    130 1124 227398860
                                                             56.5
    66 1388 0.258 0.333 0.410 107 1224 194259933 45.0
[5 rows x 28 columns]
Index(['TeamName', 'RA/G', 'DefEff', 'E', 'DP', 'W', 'L', 'W-L%', 'ERA',
        'tSho', 'H', 'ER', 'HR', 'BB', 'SO', 'R/G', 'R', 'H 1', 'RBI', 'SB',
        'SO 1', 'BA', 'OBP', 'SLG', 'GDP', 'LOB', 'salary', 'WAR'],
       dtype='object')
TeamName
RA/G
DefEff
Ε
             0
DP
             0
             0
W
             0
L
W-L%
             0
ERA
             0
tSho
             0
Н
             0
ER
             0
HR
             0
BB
             0
S0
             0
R/G
             0
R
             0
H 1
             0
RBI
             0
SB
             0
S0 1
             0
BA
             0
OBP
             0
SLG
             0
             0
GDP
             0
LOB
             0
salary
WAR
dtype: int64
Index(['TeamName', 'RA/G', 'DefEff', 'E', 'DP', 'W', 'L', 'W-L%', 'ERA',
        'tSho', 'H', 'ER', 'HR', 'BB', 'SO', 'R/G', 'R', 'H 1', 'RBI', 'SB',
        'SO 1', 'BA', 'OBP', 'SLG', 'GDP', 'LOB', 'salary', 'WAR'],
       dtype='object')
Columns to drop: ['TeamName']
RA/G
           float64
DefEff
           float64
Ε
             int64
DP
             int64
W
             int64
L
             int64
W-L%
           float64
           float64
ERA
             int64
tSho
Н
             int64
ER
             int64
HR
             int64
BB
             int64
S0
             int64
R/G
           float64
R
             int64
H 1
             int64
             int64
RBI
SB
             int64
S0 1
             int64
BA
           float64
0BP
           float64
SLG
           float64
GDP
             int64
LOB
             int64
salary
             int64
           float64
WAR
dtype: object
                                                 Correlation Matrix
                                                                                                                         1.00
         1 -0.60.250.32<mark>0.760.760.760.99</mark>0.630.810.990.71 0.6-0.420.020402070242.040300209230.01-20.160.0808005-50.3-0.140.72
             1-0.350.310.490.490.490.570.370.790.570.160.230.140.0990.1-0.10.090.0230.1-10.076.10.00-10.00-10.00-10.090.0650.49
      E -0.250.35 1 0.180.280.280.280.17-0.10.240.16.00902210.220.170.160.150.170.180.220.150.170.150.150.0840.290.3
    DP -0.320.310.18 1 0.240.240.240.32-0.20.430.320.0960.220.440.04080470.10.0305.0202050.0940.00700307.110.180.160.24
                                                                                                                        - 0.75
     W-0.7(0.490.280.24 1 -1 1-0.740.480.690.740.380.450.520.570.570.350.570.0350.170.390.590.460.0640.320.360.91
      L - 0.760.490.280.24 - 1 1 - 1 0.740.480.690.740.380.450.520.570.570.360.50.0370.170.390.60.460.0670.320.360.91
                        1 -1 1 -0.740.480.690.740.380.450.520.570.570.350.540.036.170.390.590.460.0650.320.360.91
                                                                                                                        - 0.50
   ERA -0.990.5 0.170.320.740.740.74 1 -0.610.8 1 0.73 0.6-0.420.0006009.04.00-07026.190.01-30.170.110.010.320.11-0.7
   tSho -0.6:0.37-0.1-0.20.480.480.480.61 1 0.530.6:0.460.380.270.05020501.0380.06.0340.10600408140.106.0520.209.0780.44
     H -0.810.750.240.430.690.690.690.8-0.53 1 0.8 0.380.290.620.170.107.0930.107007060365.0490.240.07010860.230.170.6
                                                                                                                        - 0.25
    ER -0.990.570.160.320.740.740.74 1 -0.610.8 1 0.72 0.6-0.40L0902.D04.894.600907.D260.20.0098.170.1D.0140.330.11-0.7
    HR -0.71-0.30600920960.380.380.380.380.380.72-0.460.380.72-1 0.40.016.260.266.00305260.020.230.01050140.366.0220.320.140.32
    SO -0.420.140.220.470.520.520.520.420.270.620.440.01-6.12 1 0.360.360.0190.360.090.150.0170.320.340.0407.220.380.51
                                                                                                                        - 0.00
   R/G -0.020409-90.1-70.04 0.570.570.5-70.090605-70.1-07.002012-60.02/07.36 1 1 0.65 1 0.0201.01 0.650.83 0.90.0850.2 0.360.65
     R-9.0270.1-0.1-6.0470.570.570.570.00090540.107.00488240.0242.36 1 1 0.64
                                                                         1-0.0109010.650.83 0.90.0830.2 0.360.65
   H 1 9.0220.1-0.150.1 0.350.360.350.040.03080903.0406004050105010501.9.650.64 1 0.640.0490.450.980.740.610.350.370.220.45
                                                                                                                        - -0.25
    RBI -0.020309-D.1-0.030.570.560.50.00407.060-D7000002-60.0207.36 1 1 0.64 1 D.0402010.640.830.90.0830.20.360.65
     SB-0.00/2902 10.1-80.02/2.03/2503/0.03/25.00/205.00/205.00/200-00/205.00/20.09/20.02/20.01/2004/9.04/2 1 -0.1/20.07/20.01-40.07-9.1-70.05-70.01-20.00/20098
  SO 1 -0.230.110.220.05-0.170.170.170.190.16.0360.20.230.290.150.0307.01-6.450.01-20.11 1 0.480.20.0760.4-0.150.13-0.2
                                                                                                                         -0.50
    BA -0.030207-50.16.09 50.3 90.3 90.3 90.0 30800 90.30 40900 9080 305 0 30901 0.650.65 0.9 80.6 40.07 30.48 1 0.770.63 0.3 90.3 30.2 20.47
   OBP -0.180.110.170.070.59-0.60.590.170.140.220.10.0140.130.320.830.830.740.830.0140.210.77 1 0.72 0.3 0.590.320.69
   SLG 9.08080030210500307.460.460.460.110.1-0.070.110.360.0310.31 0.9 0.9 0.610.910.070907 0.630.72 1 0.0709.0540.310.5
   GDP0-00-50504B.150.1D.064.06070650.0D.05020865.014.0202.090L04070805.0830.350.08-30.17-0.40.390.30.07910.0530.130.1
                                                                                                                         -0.75
   LOB --0.30.090.084.180.320.320.320.320.290.210.310.320.180.22 0.2 0.2 0.37 0.20.0540.150.330
 salary -0.14.0650.290.160.360.360.360.10.0760.170.110.140.220.380.360.360.220.360.190.130.220.320.310.130.12 1 0.35
   WAR -0.720.49-0.30.240.910.910.91-0.70.440.67-0.70.320.460.510.650.650.450.0500098.20.470.690.55
                               %T-M
                                                                                                          WAR
    Feature
                  Score
       RA/G 249.567124
0
            224.427562
4
        ERA
7
         ER
            223.365367
6
          Н
            163.631605
            100.728243
        OBP
18
12
          R
              69.618082
        R/G
              68.775512
11
        RBI
              68.428632
14
5
       tSho
              60.264219
1
     DefEff
              57.965822
10
         S0
              57.626167
9
         BB
              56.162005
19
        SLG
              34.205305
8
        HR
              33.114826
17
         BA
              32.904535
13
        H 1
              25.792274
21
        LOB
              23.716298
22
     salary
              23.508915
2
              14.901698
          Ε
16
       S0 1
               9.574437
3
         DΡ
               6.078356
20
        GDP
               1.207182
15
         SB
               0.022458
Linear Regression RMSE: 0.02352260342713923
Linear Regression R^2: 0.8697738040454528
Random Forest RMSE: 0.026306159110422445
Random Forest R^2: 0.8371294863007793
    Feature Importance
0
       RA/G
               0.474464
               0.091246
18
        0BP
4
        ERA
               0.060793
12
               0.056258
7
         ER
               0.045598
14
        RBI
               0.039433
11
        R/G
               0.039411
               0.033119
6
         Н
               0.022841
19
        SLG
10
         S0
               0.014020
13
        H 1
               0.013777
17
         BA
               0.013630
22
     salary
               0.013116
               0.013094
2
          Ε
5
       tSho
               0.010506
9
         BB
               0.010373
21
        L0B
               0.008768
         DP
               0.007941
3
20
        GDP
               0.007407
     DefEff
               0.006613
1
```



0.006197

0.005942

16

8

S0 1

HR