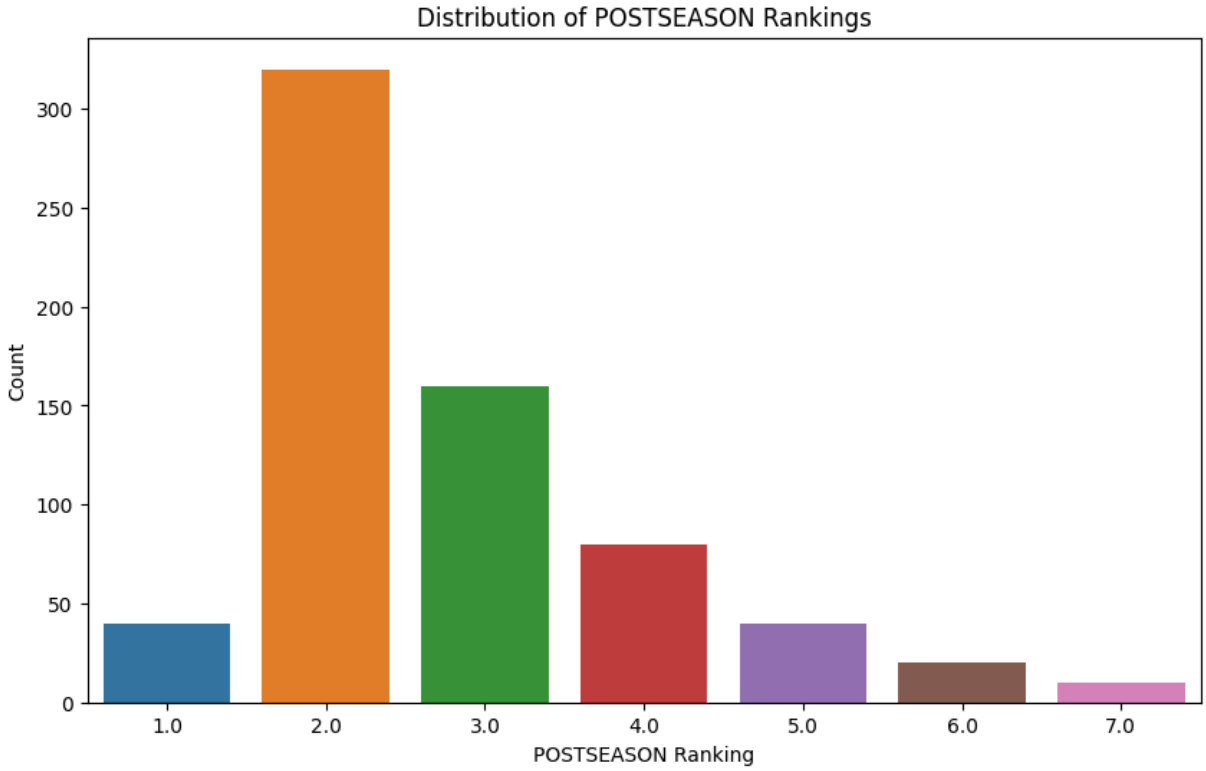
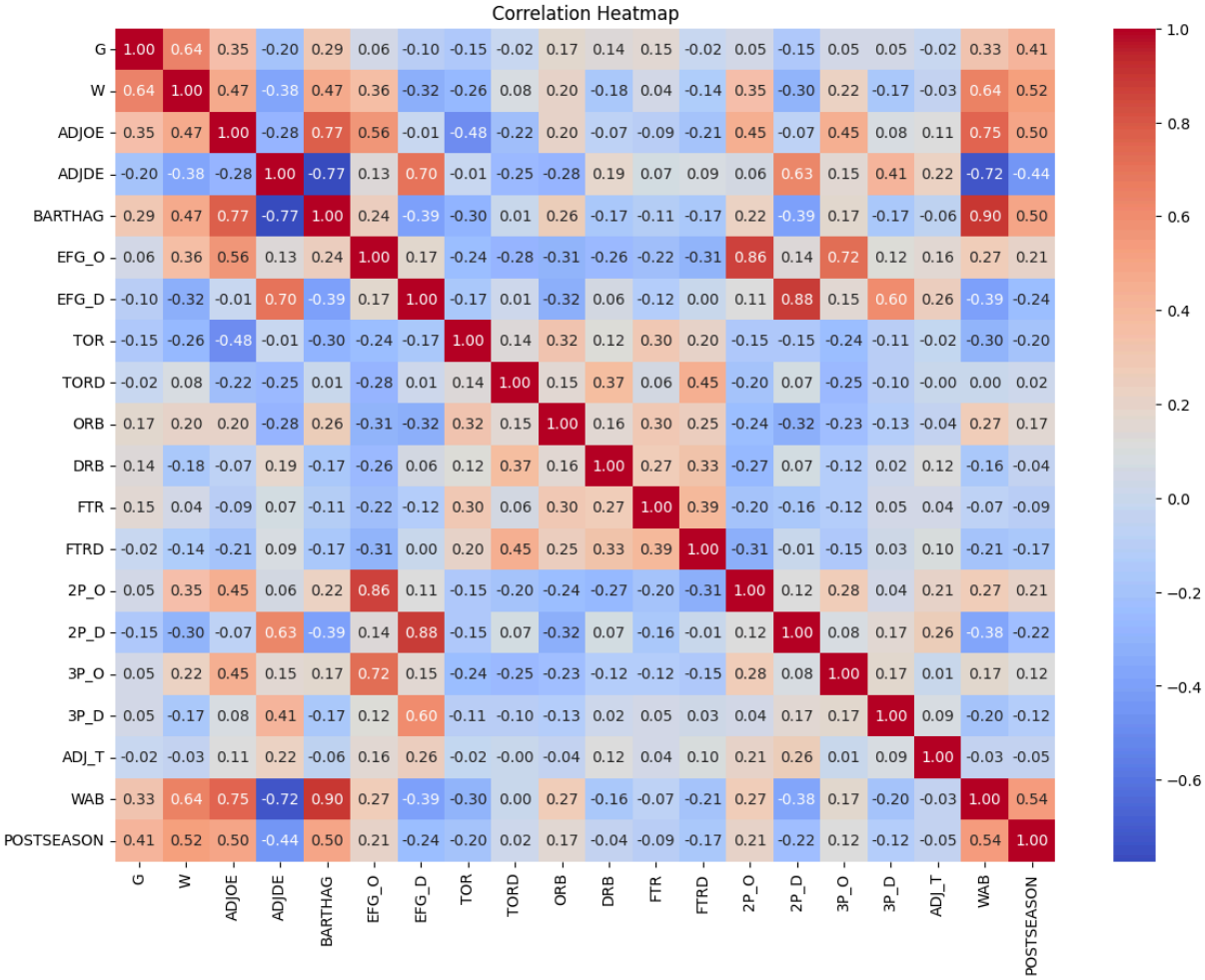
In [ ]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load the dataset
file_name = 'cbb.csv'
cbb_df = pd.read_csv(file_name)

# Display the first few rows of the dataframe
print(cbb_df.head())

# Display basic information about the dataset
print(cbb_df.info())

# Select relevant columns and drop rows with missing values
# We assume that 'POSTSEASON' indicates the ranking or outcome
cbb_df = cbb_df[['G', 'W', 'ADJOE', 'ADJDE', 'BARTHAG', 'EFG_O', 'EFG_D', 'TOR', 'T
cbb_df.dropna(subset=['POSTSEASON'], inplace=True)

# Encode 'POSTSEASON' as numerical values (e.g., 'R68', 'R64', 'R32', 'S16', 'E8',
postseason_mapping = {
    'R68': 1,
    'R64': 2,
    'R32': 3,
    'S16': 4,
    'E8': 5,
    'F4': 6,
    '2ND': 7,
    'Champ': 8
}
cbb_df['POSTSEASON'] = cbb_df['POSTSEASON'].map(postseason_mapping)

# Drop any remaining NaN values after encoding
cbb_df.dropna(subset=['POSTSEASON'], inplace=True)

# Distribution of POSTSEASON Rankings
plt.figure(figsize=(10, 6))
sns.countplot(data=cbb_df, x='POSTSEASON')
plt.title('Distribution of POSTSEASON Rankings')
plt.xlabel('POSTSEASON Ranking')
plt.ylabel('Count')
plt.show()

# Correlation Heatmap
plt.figure(figsize=(14, 10))
correlation_matrix = cbb_df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()

# Features and target variable
```

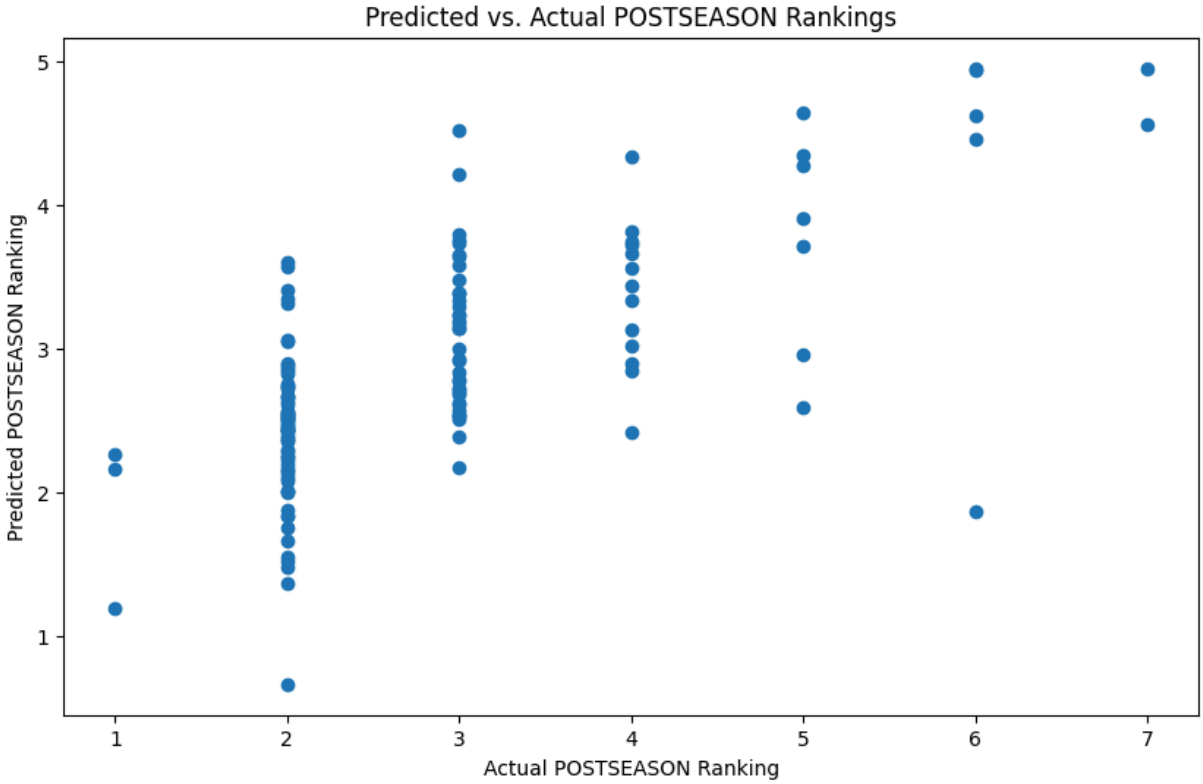
```
X = cbb_df.drop(columns=['POSTSEASON'])
y = cbb_df['POSTSEASON']

# 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_sta

# Create and train a Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
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}')

# Predicted vs. Actual POSTSEASON Rankings
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred)
plt.xlabel('Actual POSTSEASON Ranking')
plt.ylabel('Predicted POSTSEASON Ranking')
plt.title('Predicted vs. Actual POSTSEASON Rankings')
plt.show()

# Feature Importances
importance = model.coef_
feature_names = X.columns
feature_importance = pd.Series(importance, index=feature_names).sort_values(ascendi

plt.figure(figsize=(12, 8))
sns.barplot(x=feature_importance, y=feature_importance.index)
plt.title('Feature Importances')
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.show()
```

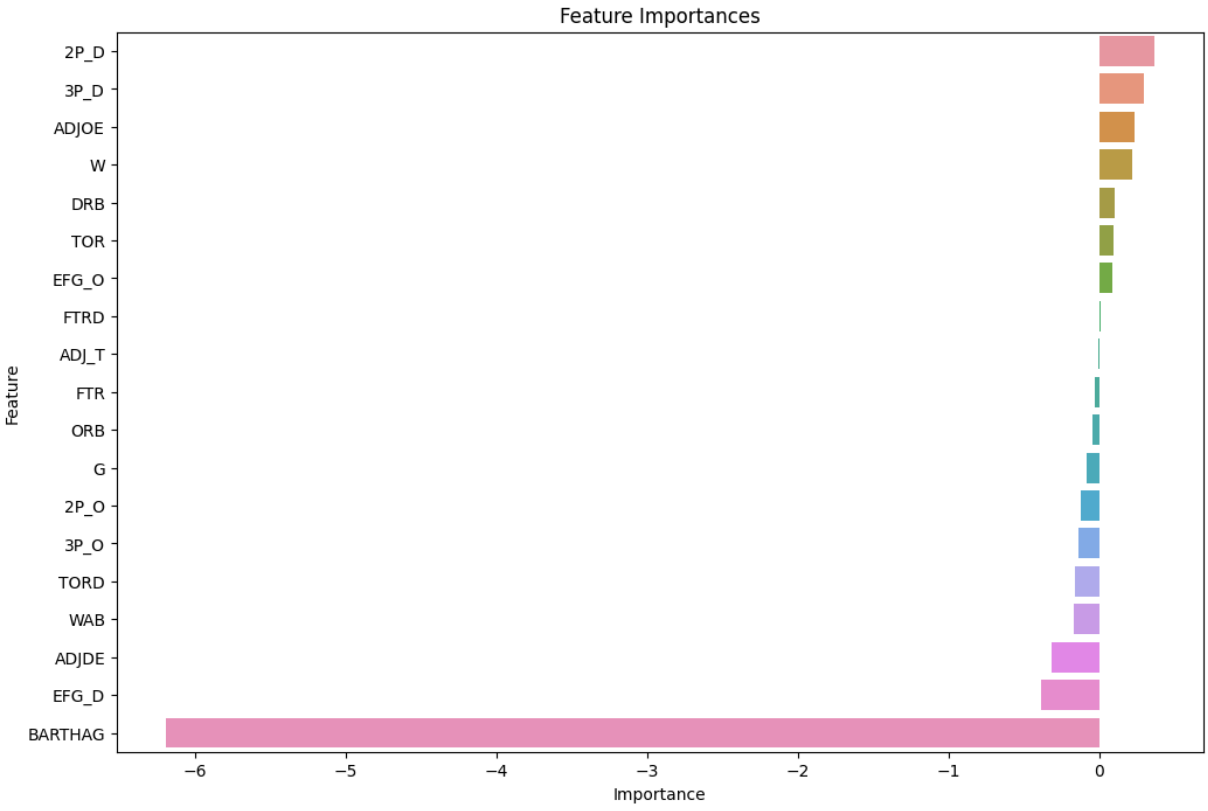
```
             TEAM CONF   G   W  ADJOE  ADJDE  BARTHAG  EFG_O  EFG_D   TOR  \
0  North Carolina  ACC  40  33  123.3   94.9   0.9531   52.6   48.1  15.4
1       Wisconsin  B10  40  36  129.1   93.6   0.9758   54.8   47.7  12.4
2        Michigan  B10  40  33  114.4   90.4   0.9375   53.9   47.7  14.0
3      Texas Tech  B12  38  31  115.2   85.2   0.9696   53.5   43.0  17.7
4         Gonzaga  WCC  39  37  117.8   86.3   0.9728   56.6   41.1  16.2

   ...  FTRD  2P_O  2P_D  3P_O  3P_D  ADJ_T   WAB  POSTSEASON  SEED  YEAR
0  ...  30.4  53.9  44.6  32.7  36.2   71.7   8.6         2ND   1.0  2016
1  ...  22.4  54.8  44.7  36.5  37.5   59.3  11.3         2ND   1.0  2015
2  ...  30.0  54.7  46.8  35.2  33.2   65.9   6.9         2ND   3.0  2018
3  ...  36.6  52.8  41.9  36.5  29.7   67.5   7.0         2ND   3.0  2019
4  ...  26.9  56.3  40.0  38.2  29.0   71.5   7.7         2ND   1.0  2017

[5 rows x 24 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3523 entries, 0 to 3522
Data columns (total 24 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   TEAM        3523 non-null   object
 1   CONF        3523 non-null   object
 2   G           3523 non-null   int64
 3   W           3523 non-null   int64
 4   ADJOE       3523 non-null   float64
 5   ADJDE       3523 non-null   float64
 6   BARTHAG     3523 non-null   float64
 7   EFG_O       3523 non-null   float64
 8   EFG_D       3523 non-null   float64
 9   TOR         3523 non-null   float64
 10  TORD        3523 non-null   float64
 11  ORB         3523 non-null   float64
 12  DRB         3523 non-null   float64
 13  FTR         3523 non-null   float64
 14  FTRD        3523 non-null   float64
 15  2P_O        3523 non-null   float64
 16  2P_D        3523 non-null   float64
 17  3P_O        3523 non-null   float64
 18  3P_D        3523 non-null   float64
 19  ADJ_T       3523 non-null   float64
 20  WAB         3523 non-null   float64
 21  POSTSEASON  680 non-null    object
 22  SEED        680 non-null    float64
 23  YEAR        3523 non-null   int64
dtypes: float64(18), int64(3), object(3)
memory usage: 660.7+ KB
None
```





Mean Squared Error: 0.7352871507639764
R-squared: 0.4878659395221894


Predicted vs. Actual POSTSEASON Rankings
Predicted POSTSEASON Ranking
Actual POSTSEASON Ranking
1
2
3
4
5
6
7


Feature Importances
Feature
Importance
2P_D
3P_D
ADJOE
W
DRB
TOR
EFG_O
FTRD
ADJ_T
FTR
ORB
G
2P_O
3P_O
TORD
WAB
ADJDE
EFG_D
BARTHAG
−6
−5
−4
−3
−2
−1
0

In [ ]: