

```
in [5]: 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('team_stats_2003_2023.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 = ['team']
for col in df.columns:
    if 'year' 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=['win_loss_perc', 'wins', 'losses', 'ties', 'points_diff', 'points', 'points_opp', 'mov', 'score_pct']) # Adjust based on your
y = df['win_loss_perc']

# 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()

year      team      wins      losses      win_loss_perc      points      \
0      2003      New England Patriots      14      2      0.875      348
1      2003      Miami Dolphins      10      6      0.625      311
2      2003      Buffalo Bills      6      10      0.375      243
3      2003      New York Jets      6      10      0.375      283
4      2003      Baltimore Ravens      10      6      0.625      391

points_opp      points_diff      mov      g      ...      rush_td      rush_yds_per_att      rush_fd      \
0      238      110      6.9      16      ...      9      3.4      91
1      261      50      3.1      16      ...      14      3.7      99
2      279      -36      -2.3      16      ...      13      3.9      96
3      299      -16      -1.0      16      ...      8      4.0      78
4      281      110      6.9      16      ...      18      4.8      115

penalties      penalties_yds      pen_fd      score_pct      turnover_pct      exp_pts_tot      \
0      111      998      26      27.9      11.3      -136.51
1      103      913      22      28.1      17.2      -177.92
2      106      891      22      21.9      17.6      -230.07
3      69      550      15      32.4      11.8      -107.89
4      126      970      23      31.8      16.6      -220.50

ties
0      NaN
1      NaN
2      NaN
3      NaN
4      NaN

[5 rows x 35 columns]
Index(['year', 'team', 'wins', 'losses', 'win_loss_perc', 'points',
      'points_opp', 'points_diff', 'mov', 'g', 'total_yards', 'plays_offense',
      'yds_per_play_offense', 'turnovers', 'fumbles_lost', 'first_down',
      'pass_cmp', 'pass_att', 'pass_yds', 'pass_td', 'pass_int',
      'pass_net_yds_per_att', 'pass_fd', 'rush_att', 'rush_yds', 'rush_td',
      'rush_yds_per_att', 'rush_fd', 'penalties', 'penalties_yds', 'pen_fd',
      'score_pct', 'turnover_pct', 'exp_pts_tot', 'ties'],
      dtype='object')

year      0
team      0
wins      0
losses     0
win_loss_perc  0
points     0
points_opp  0
points_diff  0
mov      320
g          0
total_yards  0
plays_offense  0
yds_per_play_offense  0
turnovers    0
fumbles_lost  0
first_down    0
pass_cmp      0
pass_att      0
pass_yds      0
pass_td      0
pass_int      0
pass_net_yds_per_att  0
pass_fd      0
rush_att      0
rush_yds      0
rush_td      0
rush_yds_per_att  0
rush_fd      0
penalties     0
penalties_yds  0
pen_fd      0
score_pct     0
turnover_pct  0
exp_pts_tot   0
ties         352
dtype: int64
Index(['year', 'team', 'wins', 'losses', 'win_loss_perc', 'points',
      'points_opp', 'points_diff', 'mov', 'g', 'total_yards', 'plays_offense',
      'yds_per_play_offense', 'turnovers', 'fumbles_lost', 'first_down',
      'pass_cmp', 'pass_att', 'pass_yds', 'pass_td', 'pass_int',
      'pass_net_yds_per_att', 'pass_fd', 'rush_att', 'rush_yds', 'rush_td',
      'rush_yds_per_att', 'rush_fd', 'penalties', 'penalties_yds', 'pen_fd',
      'score_pct', 'turnover_pct', 'exp_pts_tot', 'ties'],
      dtype='object')

Columns to drop: ['team', 'year']
wins      int64
losses     int64
win_loss_perc  float64
points     int64
points_opp  int64
points_diff  int64
mov      float64
g          int64
total_yards  int64
plays_offense  int64
yds_per_play_offense  float64
turnovers    int64
fumbles_lost  int64
first_down    int64
pass_cmp      int64
pass_att      int64
pass_yds      int64
pass_td      int64
pass_int      int64
pass_net_yds_per_att  float64
pass_fd      int64
rush_att      int64
rush_yds      int64
rush_td      int64
rush_yds_per_att  float64
rush_fd      int64
penalties     int64
penalties_yds  int64
pen_fd      int64
score_pct     float64
turnover_pct  float64
exp_pts_tot   float64
ties         float64
dtype: object

Correlation Matrix
wins      1.000000  0.796090  0.004050  0.205050  0.305015  0.636504  0.304030  0.483500  0.124010  0.120071  0.605050  0.091
win_loss_perc  -0.991 -1.000000  0.067022  0.402900  0.180703  0.584060  0.384039  0.481204  0.110060  0.105040  0.504032
points      0.797075  0.108860  0.090840  0.804063  0.804020  0.670040  0.602039  0.602945  0.008209  0.404070  0.058
points_diff  -0.606011  1.070010  0.100200  0.202020  0.009030  0.020010  0.161048  0.201907  0.200060  0.501610  0.003
mov      0.909088  0.110700  0.002050  0.502029  0.107014  0.504060  0.384035  0.501645  0.080421  0.705050  0.059
total_yards  0.606666  0.007314  0.036194  0.403824  0.104063  0.403050  0.293026  0.264010  0.130020  0.028150  0.139018
yds_per_play_offense  0.202020  0.021280  0.104006  1.000000  0.090909  0.596058  0.606020  0.201020  0.026308  0.062900  0.104047
turnovers    -0.490040  0.482800  0.381038  0.100994  1.070030  0.606020  0.308039  0.202020  0.382938  0.380062  0.145090  0.007
fumbles_lost  -0.290030  0.311029  0.281019  0.260922  0.711026  0.006020  0.202026  0.080080  0.284914  0.160060  0.058030  0.006
first_down    0.505010  0.005040  0.209069  0.703027  1.060040  0.797030  0.703070  0.801828  0.492440  0.102040  0.408040  0.011
pass_cmp      0.151014  0.201010  0.206040  0.106066  1.060061  0.908650  0.924080  0.401029  0.200020  0.000000  0.703040  0.041
pass_att      0.063070  0.203090  0.461040  0.600020  0.806409  1.070040  0.107050  0.502030  0.360095  0.201000  0.129057
pass_yds      0.303030  0.601040  0.200060  0.507070  0.807079  0.607410  0.701078  0.902926  0.090090  0.085029  0.601000  0.065
pass_td      0.505030  0.001040  0.400370  0.307039  0.207050  0.407710  0.207470  0.083090  0.006090  0.005020  0.703000  0.053
pass_int      0.404040  0.403040  0.303030  0.303030  0.802030  0.902030  0.102010  0.102010  0.302030  0.300090  0.203000  0.032
pass_net_yds_per_att  0.606020  0.106050  0.038020  0.902090  0.207049  0.107070  0.301060  0.101010  0.102030  0.000000  0.703000  0.063
rush_att      0.404040  0.204840  0.161020  0.200650  0.202020  0.807690  0.701068  0.202020  0.100000  0.100000  0.100000  0.042
rush_yds      0.303830  0.203820  0.160301  0.200709  0.829401  0.402020  0.150190  0.104020  0.102084  0.060090  0.090080  0.037012
rush_td      0.504840  0.601050  0.105020  0.402049  0.104029  0.202019  0.103040  0.105064  0.105070  0.080090  0.305030  0.002
rush_yds_per_att  0.121012  0.280010  0.109030  0.025029  0.102402  0.400906  0.281010  0.307051  0.400020  0.008630  0.203000  0.018
rush_fd      0.401340  0.404020  0.483028  0.303030  0.303030  0.193080  0.903020  0.302000  0.708040  0.101000  0.028400  0.039
penalties     0.121010  0.109040  0.909080  0.202020  0.000000  0.600070  0.102006  0.009050  0.000000  0.000000  1.000000  0.012
penalties_yds  0.060060  0.000000  0.402020  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.900000  0.012
score_pct     0.504040  0.700000  0.382030  0.407040  0.705040  0.805060  0.706070  0.507060  0.601020  0.483040  0.050000  1.000000
turnover_pct  0.101010  0.101010  0.101010  0.101010  0.101010  0.101010  0.101010  0.101010  0.101010  0.101010  0.101010  0.101010
exp_pts_tot   0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
ties         0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000

Feature      Score
pass_net_yds_per_att  297.659093
total_yards      264.022200
pass_td      193.548016
yds_per_play_offense  183.885058
first_down      177.420983
turnover_pct      173.368598
exp_pts_tot      164.858126
turnovers      156.743036
rush_td      151.922786
pass_int      136.712731
rush_att      102.221582
rush_yds      88.983523
pass_yds      77.741296
pass_fd      65.257359
rush_yds      56.432111
fumbles_lost      44.410812
plays_offense      38.091267
pass_cmp      15.759148
pen_fd      8.835691
penalties      5.767016
rush_yds_per_att  4.697230
penalties_yds  1.115565
pass_att      0.638217
0      0.004808
Linear Regression RMSE: 0.10724015127315736
Linear Regression R^2: 0.7052431508364485
Random Forest RMSE: 0.1208962013586984
Random Forest R^2: 0.6253943010717984

Feature      Importance
pass_net_yds_per_att  0.334012
turnover_pct      0.104472
rush_att      0.047844
pass_td      0.045416
exp_pts_tot      0.029697
rush_fd      0.023322
pen_fd      0.022422
turnovers      0.022152
penalties      0.021846
penalties_yds  0.021405
rush_yds      0.021185
rush_yds_per_att  0.020679
first_down      0.020395
pass_int      0.020163
plays_offense  0.019908
total_yards      0.019712
pass_cmp      0.018327
fumbles_lost      0.016483
pass_yds      0.014785
pass_fd      0.013955
rush_yds      0.013167
yds_per_play_offense  0.012911
0      0.006935

Feature Importance from Random Forest
pass_net_yds_per_att  0.334012
rush_att  0.104472
turnover_pct  0.104472
rush_td  0.047844
pass_td  0.045416
exp_pts_tot  0.029697
rush_fd  0.023322
pen_fd  0.022422
turnovers  0.022152
penalties  0.021846
penalties_yds  0.021405
rush_yds  0.021185
rush_yds_per_att  0.020679
first_down  0.020395
pass_int  0.020163
plays_offense  0.019908
total_yards  0.019712
pass_cmp  0.018327
fumbles_lost  0.016483
pass_yds  0.014785
pass_fd  0.013955
rush_yds  0.013167
yds_per_play_offense  0.012911
0  0.006935
```