The goal of this project is to showcase the win rate of soccer teams compared to their 'expected win rate' based on the quality of the shots each team took that game.

The dataset I am using take data from soccer games from between 2016 and 2023. The data for the games that we will be using in this project is expected goals from both sides and the number of goals from both sides.

"Expected Goals" are determined by comparing a shot taken by a player on goal to similar shots from previous games. Factors used to compare the shot to previous shots are things such as postioning of the players or speed of the shot. The percentage of the times the previous shots have gone in is how often you would expect the shot we are looking at now to go in.

For example if there is a shot taken in a game and it is similar to 20 previous shots from other games and those previous shots scored 10 times out of the 20 they were taken then you would expect the shot we are analyzing now to go in about 50% of the time. This will score a 0.5 on the expected goals scale, commonly refered to as xG.

If over the course of a game team one has an xG of 2.3 and team two has an xG of 1.5 you would expect on average that team one would win the game. Over the course of many games however the team that has more xG than the opposing team doesn't always win the game.

This project takes the data and finds the amount of times the team with the higher xG wins the game. It also looks at the margin of xG between the two teams. So we would expect a team that scores a xG a lot higher than the other team to win more often than when a team scores only slightly more xG than their opponent.

The question we aim to answer is "How often does the team with the higher score of expected goals actually win the game? Also, does the higher margin of expected goals over an opponent correlate with winning the game at a higher rate?"

In [10]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
# read in the csv
df = pd.read_csv("spi_matches.csv")
```

In [3]:
```
# find the expected goals values needed
df_cleaned = df.dropna(subset=['xg1', 'xg2', 'score1', 'score2'])
```

In [6]:
```
# create new column showing which team won the game 1 or 2
def determine_winner(team_a_score, team_b_score):
    if team_a_score > team_b_score:
        return "team1"
    elif team_b_score > team_a_score:
        return "team2"
    else:
        return "draw"
```

```
df['Winner'] = df.apply(lambda row: determine_winner(row['score1'], row['score2']), ax

print(df['Winner'])
```

```
0        team1
1        team1
2         draw
3         draw
4        team2
         ...
68908     draw
68909     draw
68910     draw
68911     draw
68912     draw
Name: Winner, Length: 68913, dtype: object
```

In [12]:
```
# find the win percentage for all games with more than 0, 0.5, 1, 2 expected goals tha
df['Expected_Score_Diff'] = df['xg1'] - df['xg2']

thresholds = [0, 0.5, 1, 1.5, 2]

def calculate_win_rate(threshold):
    filtered_games = df[abs(df['Expected_Score_Diff']) > threshold]

    total_games = len(filtered_games)
    expected_winner_wins = filtered_games.apply(
        lambda row: ((row['Expected_Score_Diff'] > 0 and row['Winner'] == 'team1') or
                     (row['Expected_Score_Diff'] < 0 and row['Winner'] == 'team2')), a

    win_rate = expected_winner_wins / total_games if total_games > 0 else 0
    return win_rate

win_rates = {threshold: calculate_win_rate(threshold) for threshold in thresholds}

print(win_rates)
```

```
{0: 0.5556726183874903, 0.5: 0.6401657715566457, 1: 0.7224405229692077, 1.5: 0.799652
2669519861, 2: 0.8534059945504087}
```

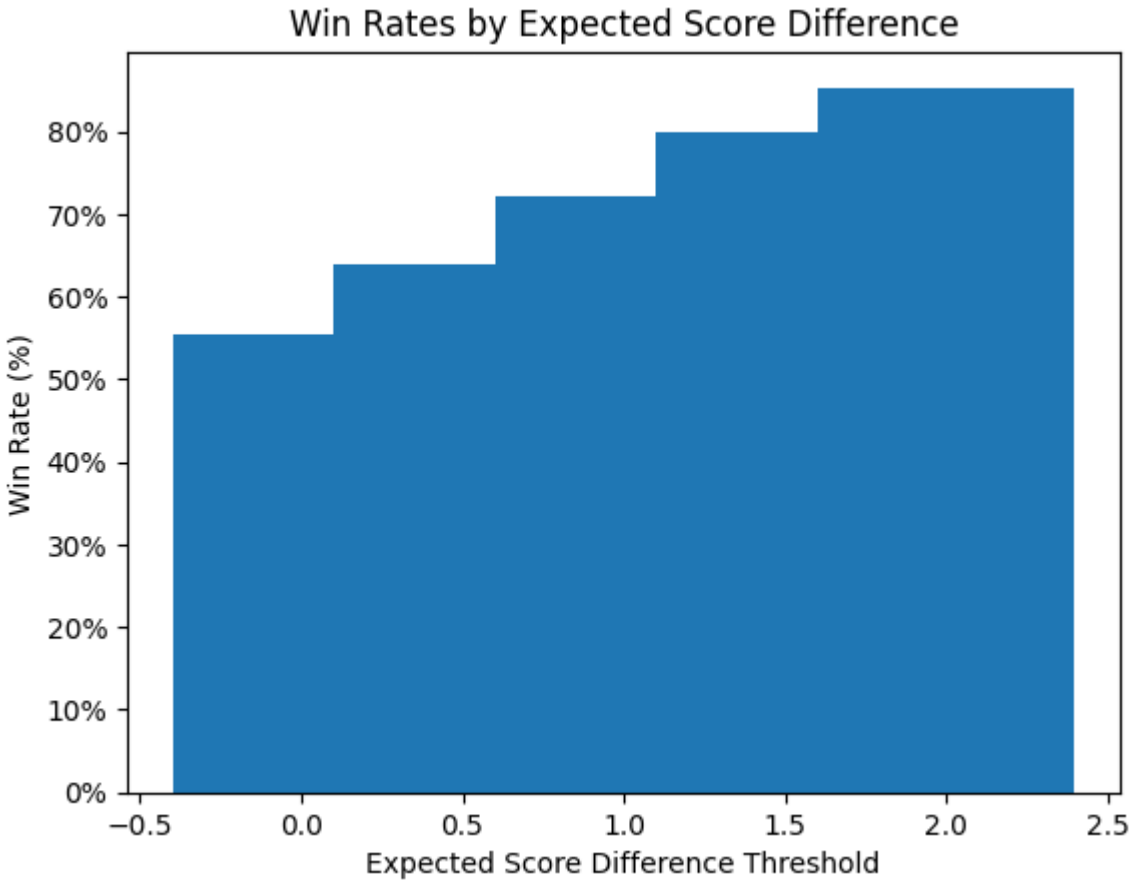
In [15]:
```
# graph the results
thresholds = list(win_rates.keys())
rates = [rate * 100 for rate in win_rates.values()]

plt.bar(thresholds, rates)

plt.xlabel('Expected Score Difference Threshold')
plt.ylabel('Win Rate (%)')
plt.title('Win Rates by Expected Score Difference')

plt.gca().yaxis.set_major_formatter(plt.matplotlib.ticker.PercentFormatter())

plt.show()
```


Win Rates by Expected Score Difference
Win Rate (%)
80%
70%
60%
50%
40%
30%
20%
10%
0%
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Expected Score Difference Threshold