Introduce the Problem

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? Does playing at home give a significant advantage to the home team? How do different leagues compare to each other in terms of theses stats?"

uce the Data

taset I am using for this project has 23 columns.

,date,league_id,league,team1,team2,spi1,spi2,prob1,prob2,probtie,proj_score1,proj_score2,importanc

: The season during which the match was played

ne date of the match (YYYY-MM-DD)

_id: A unique identifier for the league this match was played in

The name of the league this match was played in

The home team's name

The away team's name

e home team's overall SPI rating before the match

e away team's overall SPI rating before the match

The probability of the home team winning the match

The probability of the away team winning the match

: The probability of match ending in a draw (if applicable)

ore1: The number of goals we expected the home team to score

ore2: The number of goals we expected the away team to score

ance1: The importance of the match for the home team (0-100)

ance2: The importance of the match for the away team (0-100)

The number of goals scored by the home team

The number of goals scored by the away team

e number of expected goals created by the home team

e number of expected goals created by the away team

The number of non-shot expected goals created by the home team

The number of non-shot expected goals created by the away team

re1: The number of goals scored by the home team, adjusted for game state

re2: The number of goals scored by the home team, adjusted for game state

Preprocessing the Data

This data set contains much of what I need to complete the following analysis, some things that I have had to add to the data frame once I have read it in are creating some new columns. I created a new column depending on what I needed to create that specific visualization so if I needed a column that showed which team won the game I would create one by taking the score differential of the two teams and adding value, "team1", "team2", or "draw" to show the result of the game.

Data Understanding/Visualization

I created four different visualizations: Win Rates by Expected Score Difference, Actual Wins by Match Importance, Win Percentage of Higher xG Teams by League, and Home Team with Higher xG Win Rate for each League.

Storytelling

These graphs all tell one part of the story of soccer match. With all soccer games be unable to be defined by a single factor I hope to break down some of those factors in the following graphs and show a little bit about each one.

Win Rates by Expected Score Difference

This graph shows something we would likely assume without it that the team that takes more shots and shots more quality to them than their oppossing team are indeed more likely to win the match.

Actual Wins by Match Importance

This shows that across matches of all levels of importance the team with the home field advantage is more likely to win. I feel this is because important games the crowd tends to be larger and louder making the opponent team have to deal with the negative effects of that while the home team enjoys all the positives. I felt that there was a chance before I made this graph that maybe if a match was extremely important we would see both teams play more nervous and have more draws or home field advantage would be negated somewhat but that does not appear to be the case.

Win Percentage of Higher xG Teams by League

This graph is to show off how important the home field advantage is for each team. As we can see here the team with the higher expected goals does tend to win more. What is interesting here is that matches in big five leagues are having an even higher win percentage than other leagues on average in this graph. You can see that the Barclay's Premier League, German Bundesliga, Italian Serie A, and French Ligue 1 all have higher win rates than the others. Some other notable leagues like the Dutch Eeredivisie also have a high win rate in comparison to the more average leagues. I believe that if La Liga had enough games to be included on this graph they would also be among the higher half of leagues in term of win percentage. I believe this is because in the more prestigous leagues in the world it is more likely for a team to have a bigger audience a better home field advantage and a higher skill level enabling them to close out wins more often and having a lower variance in terms of losing even with higher expected goals.

Home Team with Higher xG Win Rate for each League

Very similar to the last graph this one shows much of the same in terms of leagues with higher performance in general having a higher win rate at home when they have a higher expected goals count. Differently though we see that the two highest 'leagues' here are the UEFA Europa

League and the UEFA Champions League which are different from domestic competitions because these two are tournaments featuring the best teams only from every top league in Europe. This further shows my hypothesis that the big teams tend to win more often when given the advantage. They are more solid and less likely to faulter under pressure. It is much harder to cause an upset against them with lower expected goals than other teams because of the reasons I have stated previously. Those reasons being these teams have players that make fewer mistakes, capitalize on their opponents mistakes more often, and they have larger crowds supporting them at home and distracting the opposing team.

Impact Section

I don't think this specific project will have much impact on the world. It was mainly for me to see some interesting statistics about my favorite sport and make some pretty graphs. For those soccer teams that are actually using data science at the top leagues there visualizations would probably include heatmaps of where different players score from the most or maybe a visualization showing off when the most optimal position is to play a pass or a shot in. Those teams will have far more data and specific data available than I have right now. Also there data would be more focused on the actual players of the match and have geographical locations of players while my data is more just on the outcome of different games. They would be much more focused on the specfics of each game rather than how I am showing the outcomes of a lot of different games in the big picture without really looking at the statistics of an individual game anymore than expected goals.

References https://github.com/fivethirtyeight/data/tree/master/soccer-spi

Code

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df = pd.read_csv("spi_matches.csv")
```

In [3]:

```
df_cleaned = df.dropna(subset=['xg1', 'xg2', 'score1', 'score2'])
```

In [4]:

```
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
```

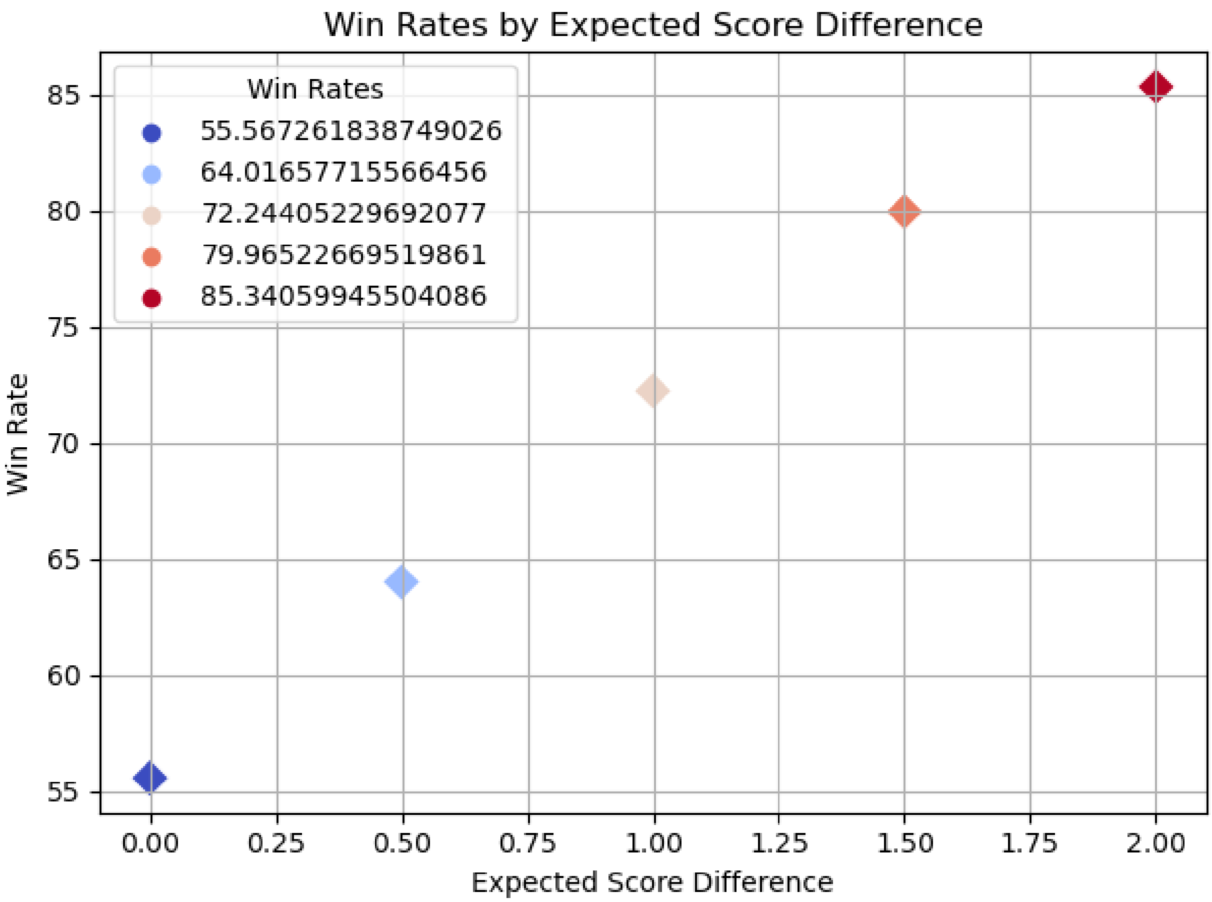
In [5]:

```python
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}
```

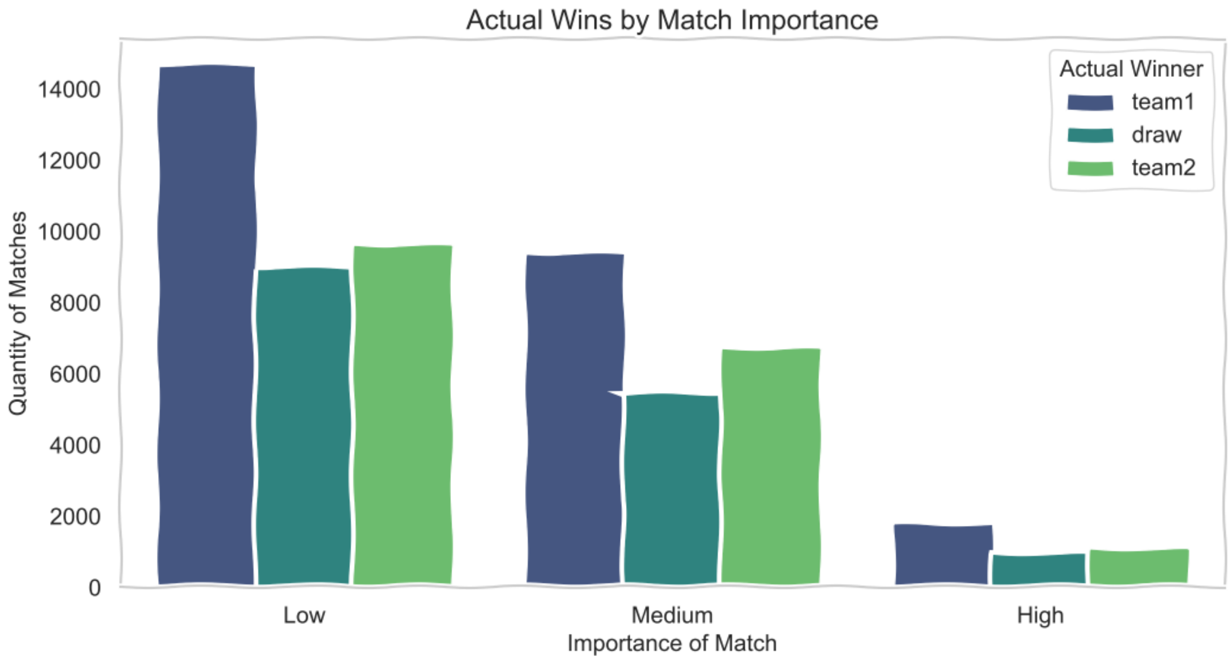
In [6]:

```python
thresholds = list(win_rates.keys())
rates = [rate * 100 for rate in win_rates.values()]

df = pd.DataFrame({'Thresholds': thresholds, 'Win Rates': rates})

sns.scatterplot(data=df, x='Thresholds', y='Win Rates', s=100,
                hue='Win Rates', palette='coolwarm',
                marker='D')

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

plt.grid()
plt.tight_layout()

plt.show()
```



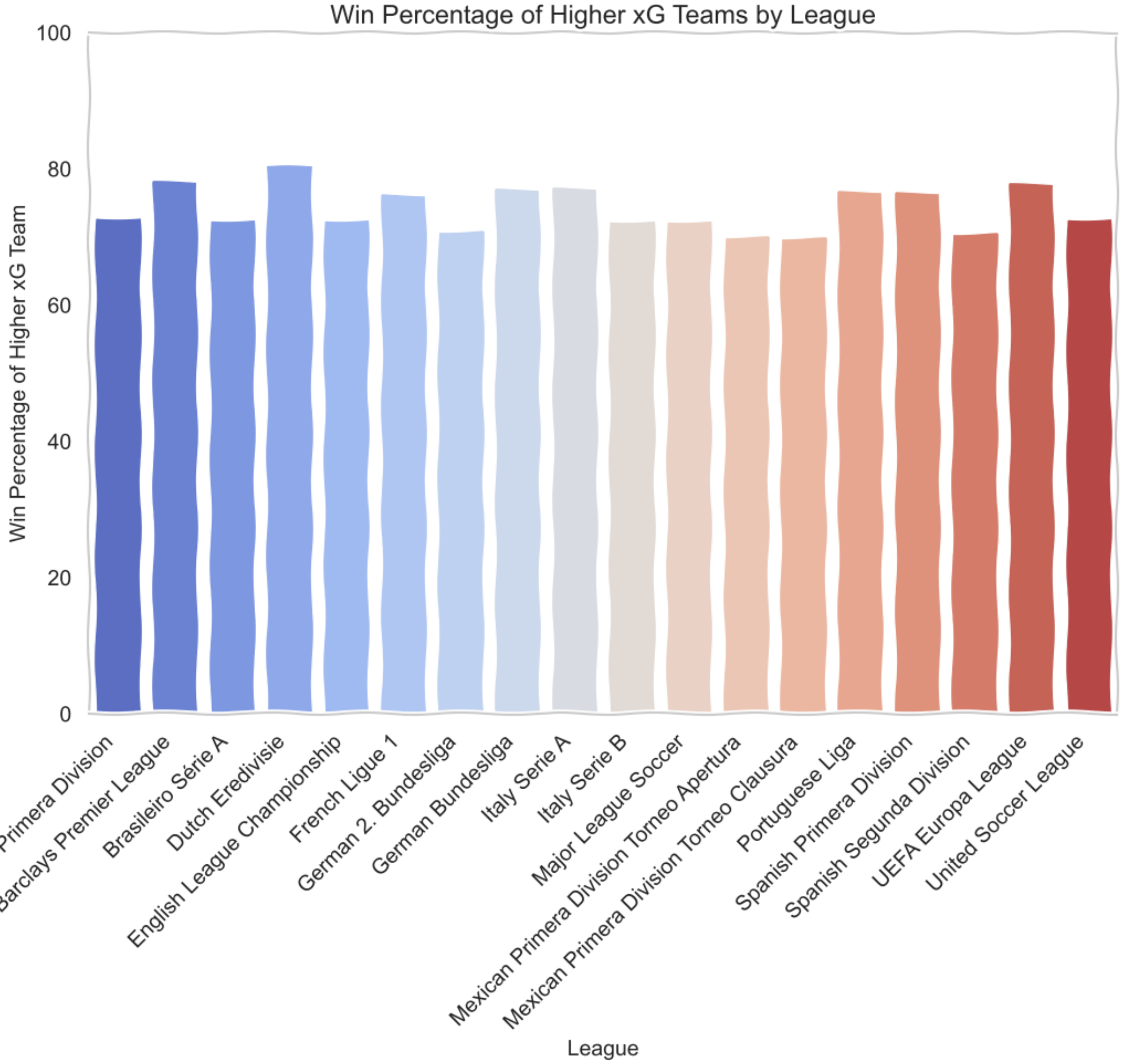
In [7]:

```
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

df = pd.read_csv("spi_matches.csv")

df['actual_winner'] = np.where(df['score1'] > df['score2'], 'team1',
                               np.where(df['score1'] < df['score2'], 'team2', 'draw'))

df['average_importance'] = (df['importance1'] + df['importance2']) / 2

df['importance'] = pd.cut(df['average_importance'], bins=[0, 33, 66, 100], labels=['Lo

plt.figure(figsize=(12, 6))
plt.xkcd()
sns.set_style("whitegrid")
sns.countplot(x='importance', hue='actual_winner', data=df, palette = "viridis")
plt.title('Actual Wins by Match Importance')
plt.xlabel('Importance of Match')
plt.ylabel('Quantity of Matches')
plt.legend(title='Actual Winner', loc='upper right')
plt.show()
```



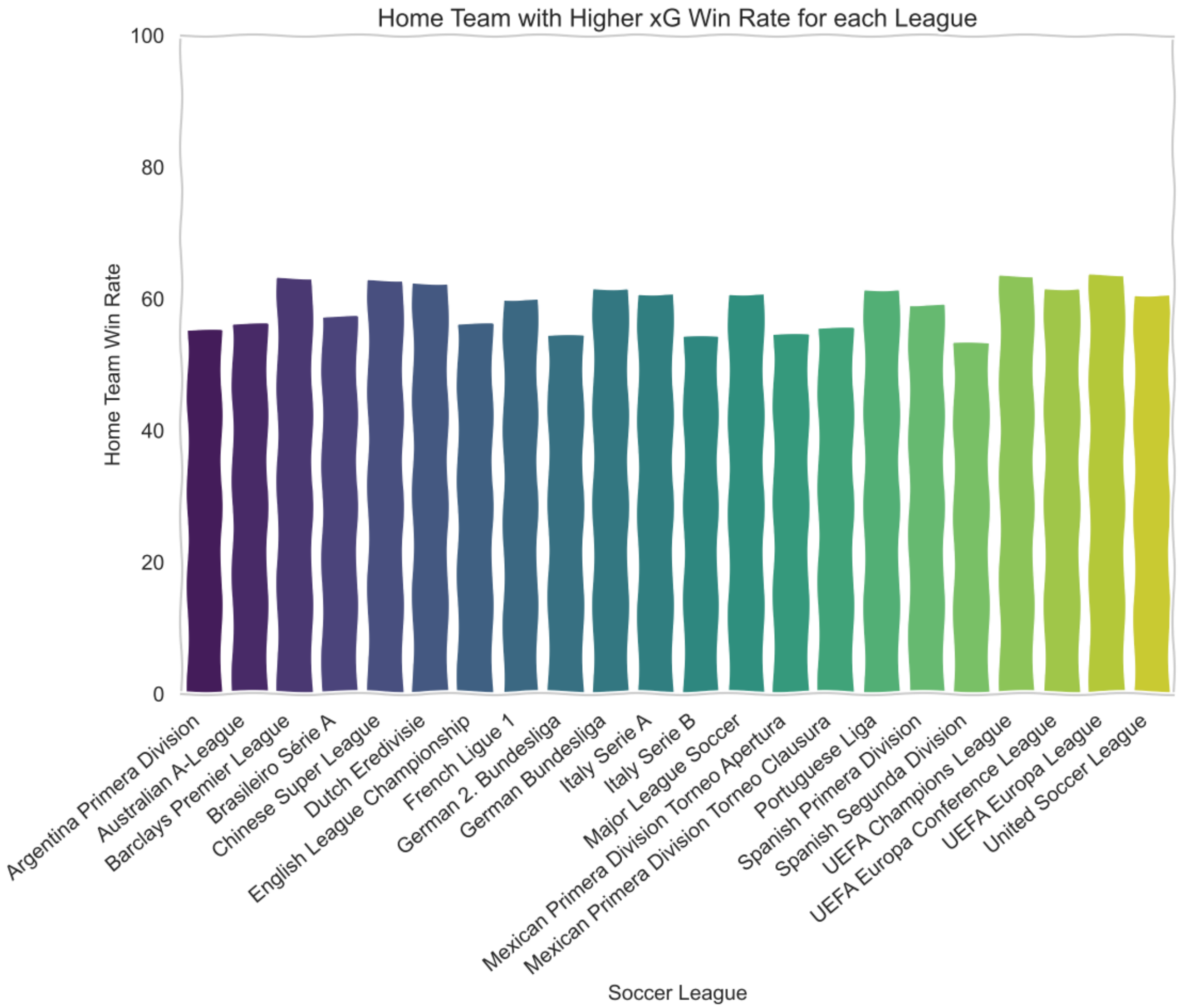
In [8]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("spi_matches.csv")

games_count_by_league = df['league'].value_counts()
leagues_with_enough_games = games_count_by_league[games_count_by_league >= 1000].index
df_filtered = df[df['league'].isin(leagues_with_enough_games)].copy()

df_filtered['higher_xg_team_won'] = None

df_filtered.loc[(df_filtered['xg1'] > df_filtered['xg2']) & (df_filtered['score1'] > c
df_filtered.loc[(df_filtered['xg2'] > df_filtered['xg1']) & (df_filtered['score2'] > c
df_filtered.loc[(df_filtered['xg1'] > df_filtered['xg2']) & (df_filtered['score1'] < c
df_filtered.loc[(df_filtered['xg2'] > df_filtered['xg1']) & (df_filtered['score2'] < c

df_filtered = df_filtered.dropna(subset=['higher_xg_team_won'])

df_filtered['higher_xg_team_won'] = df_filtered['higher_xg_team_won'].astype(bool)

win_percentage_by_league = df_filtered.groupby('league')['higher_xg_team_won'].mean().
win_percentage_by_league['higher_xg_team_won'] *= 100

plt.figure(figsize=(12, 8))
sns.barplot(data=win_percentage_by_league, x='league', y='higher_xg_team_won', palette
plt.xticks(rotation=45, ha='right')
plt.ylabel('Win Percentage of Higher xG Team')
plt.xlabel('League')
plt.title('Win Percentage of Higher xG Teams by League')
plt.ylim(0, 100)
plt.show()
```



In [9]:

```
df = pd.read_csv("spi_matches.csv")

df['home_higher_xg'] = df['xg1'] > df['xg2']

df['home_win_with_higher_xg'] = (df['home_higher_xg']) & (df['score1'] > df['score2'])

df_home_higher_xg = df[df['home_higher_xg']].copy()

home_win_rate_higher_xg_by_league = df_home_higher_xg.groupby('league', as_index=False

home_win_rate_higher_xg_by_league['home_win_with_higher_xg'] *= 100

plt.figure(figsize=(12, 8))
sns.barplot(data=home_win_rate_higher_xg_by_league, x='league', y='home_win_with_highe
#plt.grid()
plt.xticks(rotation=40, ha='right')
plt.ylabel('Home Team Win Rate')
plt.xlabel('Soccer League')
plt.title('Home Team with Higher xG Win Rate for each League')
plt.ylim(0, 100)
plt.show()
```


Home Team with Higher xG Win Rate for each League
Home Team Win Rate
100
80
60
40
20
0
Argentina Primera Division
Australian A-League
Barclays Premier League
Brasileiro Série A
Chinese Super League
Dutch Eredivisie
English League Championship
French Ligue 1
German 2. Bundesliga
German Bundesliga
Italy Serie A
Italy Serie B
Major League Soccer
Mexican Primera Division Torneo Apertura
Mexican Primera Division Torneo Clausura
Portuguese Liga
Spanish Primera Division
Spanish Segunda Division
UEFA Champions League
UEFA Europa Conference League
UEFA Europa League
United Soccer League
Soccer League