**Introduction:**

Soccer, also known as football, is a sport that brings people from all around the world together. It is played in every country and has various leagues. One of the most popular leagues is the English Premier League (EPL), which includes well known teams such as Arsenal and Chelsea. This league was founded in 1992 and it is still thriving till this day. It consists of 20 clubs and every club plays 38 games a season. There are records of these matches going back years. Our team has decided to create a database that tracks information about the EPL seasons from the last 5 years. We decided on this database because we are all interested in soccer and we believe a lot of useful analyses can be drawn from it.

Our database of English Premier League is going to consist of a variety of information. It will include data on what year the matches were played in, what clubs were involved, who was the away and home team, the amount of goals scored, the players involved in the match, the players position, the coaches names, the location the game was played at, and more. These are all the different components that make up the English Premier League and are excited to see how it all comes together.

**Target Audience:**

Since this database will provide information on teams' scores during multiple Premier League matches per season each year, it can be used for analyzing the teams and their members' progress, help make predictions for future matches, and discover interesting insights from the results. This could be especially useful for statisticians and reporters that look for patterns in past games to predict the outcome of future games between certain teams and certain players. The participating coaches, teams, players, fans, and even sponsors of the Premier League could also be utilizing this data since they’d want to know who will win, who they should support, and which teams and members they can rely on to succeed.

**Sample Data:**

Our team was able to find real data from kaggle.com’s Premier League databases which used methods such as web scraping from the websites: http://www.football-data.co.uk/ and https://www.premierleague.com/stats/clarification. Some potential data sources we plan to sort real data from are:

* <https://www.kaggle.com/zaeemnalla/premier-league> - Represents over twelve seasons of matches from 2006/2007 to 2017/2018
* <https://www.kaggle.com/prateekagrawal1405/premier-league-score-20002019> - Represents final score tables from 2000 to 2019
* <https://www.kaggle.com/mauryashubham/english-premier-league-players-dataset> - Represents individual data for each player and their value

**Potential Entities/Tables with Potential Attributes/Columns:**

1. **Matches**: a table showing general information about matches and their outcome:
   1. Match ID (integer): sequentially generated number unique to each match
   2. Date (Datetime): indicating the date when a match took place (e.g. 01/01/2001)
   3. Home club ID: home club unique ID
   4. Home (string): name of club playing home
   5. Away club ID: away club unique ID
   6. Away (string): name of club playing away
   7. Home score (integer): home club score
   8. Away score (integer): away club score
   9. Location (string): the location of the match (e.g. England, Manchester)
2. **Players**: a table of players and their club affiliation:
   1. Player ID (integer): generated sequentially to represent a unique data point for each player
   2. Club ID (integer): the unique ID of the club the player is associated with.
   3. First name (string): player’s first name
   4. Last name (string): player’s last name
   5. club (string): club membership
   6. Shirt number (integer): the number the player wears on the field
   7. Player’s position (string): the position of the player (e.g. Goalkeeper)
   8. Best player counter (integer): number of times the player won the title “Best Player” in a season
3. **Clubs**: a table to hold information about clubs:
   1. Club ID (integer): a sequentially generated number to represent a club’s unique ID number
   2. Club name (string): club’s name
   3. Coach’s first name (string): the first name of the club’s coach
   4. Coach’s last name (string): the last name of the club’s coach
   5. Number of players: number of players in the club (including substitutes)
4. **Winners:** a table to show information about clubs that won the league any season:
   1. Club ID (integer)
   2. Club name (string)
   3. Year (integer): 4-digit number representing the year the club won the league
   4. Final score (integer): their final score in the league
   5. Best player (string): if exists, the name of the player that won “Best Player” title the year the club won the league.
   6. Coach’s name (string): name of coach that led the club to win (e.g. “Smith, John”)
5. **Players and goals**: a table for players and number of goals they scored in matches:
   1. Match ID (integer)
   2. Player ID (integer)
   3. Player’s first name (string): player’s first name
   4. Player’s last name (string): player’s last name
   5. Goals (integer): if larger than zero, indicates the number of goals the player scored during the match
   6. Assists (integer): if larger than zero, indicates the number of times the player assisted in scoring a goal during the match
6. **Players and cards:** a table to count the number of cards a player received during matches:
   1. Match ID (integer)
   2. Player ID (integer)
   3. Player’s first name (string)
   4. Player’s last name (string)
   5. Year (integer)
   6. Number of yellow cards (integer): the number of yellow cards the player received during a specific season
   7. Number of red cards (integer): the number of red cards the player received during a specific season
7. **Penalties:** includes information about the penalties that occurred during a match whether because of a violation or to settle a tied match
   1. Match ID (integer)
   2. Home club ID (integer)
   3. Away club ID (integer)
   4. Penalty kicker ID (integer): penalty kicker player ID
   5. Goalkeeper ID (integer): goalkeeper player ID
   6. Scored (boolean): whether penalty kicker scored the goal or not

**Entities/Tables We Will Not Include in the Database:**

The goal of our database is to allow users to examine the progress and changes of a player or team within the Premier League throughout the last 5 years in order to gain insight or make predictions. Therefore our database would exclude information such as:

* A table about clubs not affiliated with the Premier League would not be included in our database. Our database focuses on Premier League teams to appeal toward the viewers of the Premier League.
* A table about matches that are five or more years old in order to keep our database up to date and relevant. We want to limit the age of the matches that are included in our database in order to allow users to have more relevant analysis.
* A table about the number of viewers would not be included in our database. Since our primary focus is on the progresses and changes of players and teams information involving the audience and viewers would be not directly correlated to the players and teams.
* A table of the sponsors would also not be included since our database’s primary focus is on the changes and progression of the teams and players.
* Our database will not include any personal information about the players such as:
  + Family Members
  + Addresses
  + Numbers
  + Social Media Platforms

**Questions That the Database Will Be Able To Answer:**

1. Which players with yellow cards scored more goals during a season compared to seasons when they had less yellow cards?

2. What location did a particular club win at the most?

3. In different matches, how many assists did a player have in a certain position versus another?

4. What is the average number of goals a player scored in a particular time frame versus another?

5. What is the most common number of goals scored at a particular location?

6. What is the average number of goals for a club’s players when the club was playing at home versus when playing away?

7. Which locations had less than a certain amount of total goals scored by a particular team over a given time frame?

8. How many clubs had more than x matches where the final score of the opposing team was zero?

9. Which players (from a specific club) with more than x red cards during a season scored a certain number of goals?

10. Out of clubs who won the league, which players scored the most goals in the last match played?

11. Over a certain period, which team with a penalty rate less than x had the most draws total?