Data Description (bem74 and tst42):

- What are the observations (rows) and the attributes (columns)?

Each row in both of our datasets represent one player

Soccer Dataset:

The listed attributes of each player are:

1. League: What soccer league the player plays in
2. Team: What team the player is currently on
3. Name: First and last name
4. Position: The players position (Goalkeeper, Defender, Midfielder, or Forward)
5. Age: Player age in years
6. Height: Player height in inches
7. Weight: Player weight in pounds
8. BMI: Body Mass index of player, calculated by inputting height and weight into the following formula: BMI = 703 x weight / $[height]^2$

2019 Baseball Dataset:

The listed attributes of each player are:

1. Team Abbreviation
2. Name
3. Age
4. Bats (R or L)
5. Throws (R or L)
6. Height
7. Weight
8. Years in League
9. Games Played
10. Games at Positions (P, C, 1B, 2B, 3B, SS, LF, CF, RF)
    a. This is because many baseball players play multiple positions
    b. Analysis will likely narrow down position data for comparison sake
11. WAR (A very standardized baseball player value metric)
12. Salary
13. BMI: Body Mass index of player, calculated by inputting height and weight into the following formula: BMI = 703 x weight / $[height]^2$

- Why was this dataset created?
    - This dataset was created to examine what body types lend themselves to soccer vs. baseball. As the project expands, we will develop more specific questions, and likely scrape more data from the internet. We may also add another sport, like basketball for example, to the examination.
- Who funded the creation of the dataset?
    - Teo (tst42) was responsible for the scraping and creation of the soccer data while Bryce (bem74) was responsible for the MLB data. We both contributed to the BMI and height data generation and cleaning.
- What processes might have influenced what data was observed and recorded and what was not?
    - With the MLB data, the table that was scraped included the relevant BMI, age and position data we were looking for that could be compared across sports and leagues. The table also included WAR, years in league and salary which we decided to include should it be relevant to a later analysis. Some salary data is missing on players making league minimums or if salary data just wasn't available for certain players. If we use salary in a later analysis, players lacking salary data would be omitted.
    - An important data consideration in baseball is that players play multiple positions hence having a data column for each position played. Analysis will likely entail standardizing this data to fit one position or position category (i.e. Infield).
    - Soccer data was scraped and observed based on its relevance to our project.Performance related datapoints (goals scored, etc) were not scraped. If in the future we would like to expand our examination to include how good each player is, I will research lists of the best players in each league and make a binary column called "Top 100 in league?" or something similar.
- What preprocessing was done, and how did the data come to be in the form that you are using?
    - The data was scraped from the following websites:
        - https://www.mlssoccer.com/players
        - https://scores.nbcsports.com/epl/rosters.asp
        - https://www.baseball-reference.com/teams/ARI/2019-roster.shtml
            - This is one of the 30 MLB teams 40-man roster (used 30 sites)
    - After scraping data related to our project from those websites, we needed to clean the data (remove rows with empty cells, organize the rows in a logical order, process height and weight into integers (They came in 6'3 form and 6-3 form))
    - After data cleaning, we created a method to calculate BMI, and applied it to our datasets, adding a new attribute for each player.
    - Finally, we joined the datasets of the two soccer leagues into one soccer player dataset. We added a 'league' row to keep track of where each player plays.
- If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?
    - People were not involved, all data was taken from the internet.

- Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted in a Cornell Google Drive or Cornell Box).
    - Html files which were scraped from can be found at this link: https://github.com/bemphis/2950-Final-Project-bem74-tst42