Data Collection Report

Kiran Gadhave - u1143683
Matthew Bradley - u1206298

February 13, 2019

- **Source:**
  The data is taken from a Kaggle repository located at following url:

  $$https://www.kaggle.com/hugomathien/soccer/version/10$$

  The data is in form of an sqlite database which we downloaded and processed as described below.

- **Size:**
  The original sqlite database has multiple tables which have varying sizes. The table below describes the dimensions of each:

Table	**Dimensions (Rows × Columns)**	**File Size (Mb)**
Country	11 × 2	< 1
League	11 × 3	< 1
Match	25979 × 115	279.4
Player	11060 × 7	< 1
Player Attributes	183978 × 42	30.1
Team	299 × 5	< 1
Team Attributes	1458 × 2	< 1

  Of these tables the most important for us in our analysis is the Player attributes table.

- **Format:**
  For the Player_attributes dataset, each player has multiple statistics about him like heading accuracy, crossing accuracy, dribbling, etc. The players can be modelled as vectors using this data. We plan on using clustering techniques to group the players into different play styles according to their stats. All the stats are ratings on scale of $1 - 100$ so they are compatible and we can use various distance functions taught in the class.

- **Preprocessing:**
  The data was in a sqlite database and we saved it one csv file per table. The data is clean for all tables except Match table. which has missing values for player positions in some matches. We don't think this would hamper our analysis because we dont plan on using the Player positions yet.

- **Simulation:**
  We can easily simulate a player by randomly drawing from a number in range of $1 - 100$ for each stat. A vector of such number can represent a player.