MAJOR DATABASE DESIGN DECISIONS

We received 5 tables but decided to only move forward on only 4 that is between 2003 and 2019 seasons. The major variables for this projects are player_id, team_id and game_id. Each of these variables are integers serial numbers. Creating these databases was a little bit difficult and some decisions had to be made which may or may not affect the accuracy of this report. There was minimal preprocessing involved in this effort. First we separated names into first name and last name for the sake of normalization. Additionally, there were typos on the dates and thus we had to modify them as well. Barely any missing data and thus minimal effort on that front. Regardless, it is important to share with you our decisions in creating this database should you need references. The first major decision was to split the original database of “players_stats” because there were merging issues with the other original player table because we did not feel confident the player”s id related to each other. The second main decision is to split one of the original database into what is now called “final_away_team_numbers” and “final_home_team_numbers” because we thought it would be easier for others to perform analysis on the basis if the player or team is away or home. Next, we name each foreign key differently for potential ease of use when writing sql statements and having to join tables. We unfortunately could not import any information on coaches because of the level of inaccuracy involved. For example, in one of the tables there was coaches information however for one point . Thus, we could ot say that the coaches listed were the actual coaches for the same team between 2003 and 2019.