Stage 1

Group 24: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

NBA Database

The dataset we have chosen is the NBA database. The dataset has 16 csv files that contain data relevant to each file. Each of the 16 csv files has a large number of records and attributes (too many to list). The game, game_info, game_summary files all contain data relevant to games such as attendance, box scores, dates etc. The common_player_info, draft_combine_stats, and draft_history files all contain data relevant to actual players such as name, experience, school, height, etc. The team, team_details and team_history files contain data respective to each team such as founding year and year active till. The other_stats and play_by_play files contain more niche data such as the largest lead for a particular team in a certain game and other similar stats.

The data is not ready to be inserted into a database and cleaning is necessary. Each csv file must be thoroughly inspected to identify redundancy or empty values. If the attributes that have empty values are deemed important/interesting, further steps will be taken to populate those values. That can be through the addition of another dataset. We must identify which files are necessary for our database and how big the file size is. Some of these files contain tens of thousands of records so to cut down on that, we will choose a cutoff date and work with that. Another aspect we need to address is to ensure that the data format is more suitable to implement into our database. For example, we can get rid of leading zeroes on some fields and then remove the presence of unwanted special characters. The method of how we will be cleaning the data is using R, RStudio, and excel, to trim and clean the data.

We are planning on finding a dataset that lists team rosters for our date range. This is to connect teams in a particular season to players in a particular season. We will also add a dataset of players stats throughout the season range we have chosen. These new

datasets will be connected to our current dataset since we have information on players and teams. The addition of these new statistics will supplement our current data and will result in potentially interesting queries later. We are expecting 12-14 tables and over 20000 rows. Other cleaning processes or datasets may also be consulted if feedback states our current methodology is insufficient.