Stage 3 Data Summary

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

Links to datasets:

1. NBA Database
2. https://www.kaggle.com/datasets/justinas/nba-players-data,
3. https://www.kaggle.com/datasets/szymonjwiak/nba-traditional,
4. https://www.basketball-reference.com/leagues/NBA_2017_coaches.html (for this link, we get coaches information all the way up to the 2021-22 season, which you can find by clicking the button in the top left corner to change seasons)

The first 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 that vary (some with over 10000 records in total) and attributes (too many to list). The reason why the number of records in each file are not the same and why some may have more or less records can be attributed to the inherent connections between each file and what they actually represent. For example, the files corresponding to players will be less than the files referring to games played (since each game is a distinct record and players can participate in different games). The game, game_info.csv, game_summary.csv files all contain data relevant to games such as attendance, box scores, dates etc. The common_player_info.csv, draft_combine_stats.csv, and draft_history.csv files all contain data relevant to actual players such as name, experience, school, height, etc. The team.csv, team_details.csv, and team_history.csv files contain data respective to each team such as founding year and year active till. The other_stats .csv and play_by_play.csv files contain more niche data such as the largest lead for a particular team in a certain game and other similar stats.

The 2nd dataset we have chosen contains basic information on the players including demographic information such as age and weight, biographical information like draft year, and simple stats for career averages. The dataset contains 12.8 thousand rows of data. The primary reason we have added this dataset in addition to the first one we found, is so that we can find season-specific information on what teams players have played on and for what season. This dataset is linked to all of our other datasets since we have player-specific information and team information as well. The dataset requires us to remove players to who don't fit our timeline of seasons that we want to cover (i.e. 2016-2017 to 2022/2023). Thus we need to remove players who have played before and after this date range, as well removing attributes that we already have (like career averages since we have another way of dealing with that).

The third dataset contains team stats and player stats for individual games across numerous seasons. It is linked to all of our other datasets since we have player information and the teams they played on, along with the seasons they played on them. Similar to the last dataset, we have to remove seasons that aren't in our time span. It contains attributes like points, rebounds, and steals, and it has 68.2 thousand rows. The last dataset contains coaches information throughought multiple seasons and what teams they coaches. This is connected to our other datasets since coaches coach teams and we have team information (and the players that played for them) in each season.

There are numerous entities that are contained in our dataset. For instance, we have entities like players, coaches, officials, games, and many more. One benefit of this dataset is that these entities form a tight connection with one another, resulting in a graph that stays connected even when removing other tables/connections. In other words, relationships exist with a wide variety of entities, rather than only stemming from 1 or 2. To illustrate this wide range of entity connections, some examples include:

- Players Have Stats
- Referees Officiating Games
- Teams PlayIn Stadium
- Many more!

The data is not ready to be inserted into a database and cleaning is necessary. Our first step will be to thoroughly inspect each csv file 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 such as merging other datasets containing the missing information, otherwise the values can be safely discarded. Next, we will identify which files are necessary for our database. We will consider what entities and relationships each file contributes to guide our decision making in keeping/removing certain files. For example, play_by_play.csv is a file we are already dropped since we're aiming for a broader scope, rather than focusing on the moment-to-moment instances within a specific game. Once we are confident that each field is populated with values, we will validate the format of these values and alter them for easier use if necessary. This can be removing leading zeroes and removing the presence of unwanted special characters. After this point, our data should be ready to use (although future changes/additions may be made based on the feedback received for this part). Some of these files contain tens of thousands of records with some missing fields. To cut down on that and maintain the "interestingness" of our dataset (by not removing these fields that we want to keep), we will choose a cutoff date and work with that. The method of how we will be cleaning the data is using R, RStudio, and excel, to trim and clean the data. R will primarily be used to handle larger amounts of data and more tedious tasks. This can be in the form of checking which values in each file are null or empty, trimming entire columns that we have deemed as unnecessary, or removing rows that fall under some criteria (e.g. games that took place before the 1970s). We all have experience in utilizing R to analyze and clean data which is the main driver for choosing it for this component of our project. For simpler processes that don't require the intricate tools provided by R, we may also use Excel for our cleaning.

Note: the link provided above to our dataset will show that the csv files have a total size greater than 300 MB. However, the majority of this size is due to the play_by_play.csv file. We are not planning to use this file. Thus, the total size of our dataset is no larger than 300 MB, which adheres to the project guidelines.

We have over 12 tables and across all tables, we'll have around 200000 rows.