Summary of Dataset

For our project, we have chosen to use data from the 2024-2025 NBA season. The dataset includes game-by-game stats on every player, team data, player data, draft data, combine data, schedule data, and coach data. The complete dataset is drawn from eight different files with a total of 36092 rows and 151 columns. The size of the dataset is a great fit for this project as it is large and diverse enough to be broken down into 10 or more tables, while still being manageable and complex enough to make some interesting queries. In the later stages we will trim attributes that aren't relevant to our needs.

As mentioned, there are 8 different files that this database will be taking information from. The `common_player_info.csv` file contains 4,171 rows with 33 attributes describing player details, including `person_id`, `height`, `weight`, `birthdate`, `school`, and `position`. The `draft_combine_stats.csv` file has 1,202 rows and 47 attributes of player measurements and combine stats from the NBA Draft Combine, such as `player_id`, `season`, `standing_vertical_leap`, `max_vertical_leap`, and `bench_press`. The `draft_history.csv` file has draft outcomes for players from 1947–2023, with 7,990 rows and 14 attributes, including `person_id`, `season`, `round_number`, `round_pick`, and `overall_pick`. The `player.csv` file holds a simpler set of player records from the 1947-2023, that contains 4,831 rows with 5 attributes, including `id`, `first_name`, `last_name`, `full_name`, and `is_active`. The team.csv file contains 30 rows representing all NBA teams, with 7 attributes such as `id`, `full_name`, `abbreviation`, `city`, `state`, and `year_founded`. These csv files were taken from https://www.kaggle.com/datasets/wyattowalsh/basketball.

The `database_24_25.csv` contains detailed game-by-game stats for every player in the 2024-2025 season, with 16512 rows, and 25 attributes. Some attributes include, `Player`, `Tm` (team), `Opp` (oppponent), `Res` (result), `MP` (minutes played), `FG` field goals made, `FGA` field goal attemps, `ORB` offensive rebounds, and many other performance metrics. This csv file was taken from https://www.kaggle.com/datasets/eduardopalmieri/nba-player-stats-season-2425.

The `coaches_24_25.csv` contains 35 rows and 22 attributes that summarize the 2024-2025 coaches. Attributes include `Coach` (name), `Tm` (team), `Seasons_with_franchise`, `Seasons_overall`, and attributes about coaching performance in the regular season and playoffs, including games coached, games won, and games lost. This csv file was taken from https://www.basketball-reference.com/leagues/NBA_2025_coaches.html.

Lastly, the `NBA_24_25_schedule.csv` file records the season schedule with 1321 rows and 12 attributes, such as `date`, `Start` (time start ET), `Vistor`, `Home`, `Away_PTS`, `Home_PTS`, an other attributes such as attendance and arena name. This csv file was taken from https://www.basketball-reference.com/leagues/NBA_2025_games.html. The pandas library was used to read each months schedule into one csv file.

Cleaning plan

Most of the csv files are clean and formatted correctly. However, we will have missing values in most of the attributes for rookies who were drafted in 2024 in the `draft_combine_stats.csv`, `draft_history.csv`, and `common_player_info.csv` tables. This is because those files only include information on players who were drafted between 1947 and 2023. This affects a small portion of the dataset, as according to Basketball-Reference, only 53 out of the 569 players were drafted in 2024. We will initially load these players missing fields as NULL. If time permits, we'll manually fill the fields missing in `draft_combine_stats.csv`, `draft_history.csv`, and `common_player_info.csv` for players drafted in 2024 using public sources.

Next, we will need to fill the `player.csv` file with the players drafted in 2024. To do this we will create a list of all unqiue players from the `players.csv` and `database_24_25.csv`. The players in `database_24_25.csv` that are not in `players.csv` will be the players drafted in 2024. We will then insert these players into `players.csv` and `common_player_info.csv`, leaving all missing values as NULL. The `Player` attribute in `database_24_25.csv` which contains the players full name, will be split into the players first and last name to align with the attributes of `players.csv`.

Since `common_player_info.csv`, `draft_combine_stats.csv`, `draft_history.csv`, and `player.csv` contain information on players dating back to 1947, we will filter for players active in the 2024-2025 season.

Lastly, `Coaches_24_25.csv` has some blank cells, specifcally the columns for playoffs. This is because some coaches did not make it to the playoffs, Thus for these values we will fill it with 0 or null.

Connectedness

The `player.csv` file links to `team.csv` through the `team_id` attribute. `player.csv` also links to `common_player_info.csv`, `draft_history.csv`, and `draft_combine_stats.csv`, through the `player_id`. The `database_24_25.csv` file links to `player.csv` through the `full_name` attribute in `player.csv` and the `Player` attribute in `database_24_25.csv`. `Coaches_24_25.csv` links to `team.csv` through the `Tm` attribute in `Coaches_24_25.csv` and the `abbreviation` attribute in `team.csv`. Lastly, `NBA_24_25_schedule.csv` can be linked to `team.csv` through the `full_name` attribute in `team.csv` and either the `Home` or `Vistor` attributes in `NBA_24_25_schedule.csv`. This makes all data sets fully connected, so we can get from a dataset to any other dataset through the connected graph. Once the database is set up most, if not all tables will give ID's to each record, so we don't need to rely on the name of something as the unique identifier. The current set up of name being the primary/foreign key is so display how the tables are connected.

Links

- Link to `common_player_info.csv`, `draft_history.csv`, `draft_combine_stats.csv`, `player.csv` , and `team.csv` https://www.kaggle.com/datasets/wyattowalsh/basketball
- Link to `coaches_24_25.csv` https://www.kaggle.com/datasets/eduardopalmieri/nba-player-stats-season-2425
- Link to `NBA_24_25_schedule.csv` https://www.basketball-reference.com/leagues/NBA_2025_games.html.
- Link to `database_24_25.csv` https://www.kaggle.com/datasets/eduardopalmieri/nba-player-stats-season-2425

Project Timeline

Note

For roles being an organizer will mean that you handle a good amount of the work on that stage. The assists are there to provide feedback, add smaller parts and help review before finalizing the stage

Reminder that dates are variable here, setting these dates is going to net stage completion before the due date without going over.

This will be submitted with the first 2 Stages planned out

Stage 1 Data Discovery - September 26th

Organizer: Johnny

Assists: Kameron, Dylan

- Timeline of stage 1-2 done by September 20th **FINISHED**
- Rough copy of summary of dataset should be done by September 24th **FINISHED**
- Review phase of written summary and chosen data by September 25th

Stage 2 ER/EER Diagram (Optional) - October 3rd

Organizer: Dylan

Assists: Johnny, Kameron

- 1 paragraph reminder of chosen data from stage 1 feedback should be done by September 28th (or as soon as feedback is given)
- Rough draft of ER diagram with point form notes should be done by October 1st
- Review phase complete after October 1st

Stage 3 Database Design - October 10th

Organizer: Dylan

Assists: Johnny, Kameron

- 3-5 paragraph summary of data October 4th
- ER/EER diagram with justification October 4th
- Review the diagram October 4th
- The final relational model with description of transition from ER/EER to relational (merging and normalizing) October 7th
- Final review October 10th

Reflection #1 - October 10th

Everyone participates

- Get 3-5 paragraph rough copy of the reflection done by Oct 7th
- Do a review of the reflection by Oct 9th

Stage 4 Query check-in (optional) - October 24th

Organizer: Kameron

Assists: Johnny, Dylan

Stage 5 Query Design - November 7th

Organizer: Kameron

Assists: Johnny, Dylan

Stage 6 Interface Design - November 21st

Organizer: Johnny

Assists: Kameron, Dylan

Reflection 2 - November 21st

Project Demo - December 1st-5th

Final Report - December 5th

Reflection 3 - December 5th