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 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.

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 contemplating about dropping 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 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 more 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.

Project Timeline

▼ Part A: Designing a Database

▼ Stage 2) ER Diagram (optional submission) → DUE ON OCTOBER 3

Task: 1 paragraph reminder of chosen data based on feedback from stage 1

Description: A paragraph defining and explaining the data we have chosen. If any feedback on stage 1 was provided, it will be applied to the dataset choice and mentioned in this paragraph.

Due Date: Oct 1, 2025

Assigned to: All members will work together on this (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: ER Diagram

Description: All members will individually complete a rough draft for the ER Diagram. This is so that we can see all the different ways we interpret the data. The three ER Diagrams will then be used as a guide to create one final diagram.

Due Date: Oct 1, 2025

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

Task: Justifications for ER Relationships

Description: Point form notes that will provide reasoning for the relationship choices.

Due Date: Oct 2, 2025

Assigned to: Guwinder Khandal

Task: Justifications for ER Cardinality

Description: Point form notes that will provide reasoning for the cardinality choices.

Due Date: Oct 2, 2025

Assigned to: Aaron Herrera

Task: Justifications for ER Participation Constraints

Description: Point form notes that will provide reasoning for the participation constraints

Due Date: Oct 2, 2025

Assigned to: Jerome De Guzman

▼ Stage 3) Database Design (8%) → DUE ON OCTOBER 10

Task: 3-5 Paragraph summary of data, updated from stage 1 as appropriate

Description: Short description of what the data is and how much there is.

Due Date: Oct 7, 2025

Assigned to: Gurwinder Khandal

Task: Reviewing data summary paragraph

Description: Review the original paragraph to ensure it fully captures the dataset

Due Date: Oct 8, 2025

Assigned to: Aaron Herrera and Jerome De Guzman

Task: ER/EER Diagram

Description: A final ER/EER diagram will be made with participation and cardinality constraints. Supporting text will also be written up to justify the design choices.

Due Date: Oct 8, 2025

Assigned to:

- **Task: Justifications for ER Relationships**

  Description: Point form notes that will provide reasoning for the relationship choices.

  Due Date: Oct 8, 2025

  Assigned to: Jerome De Guzman

- **Task: Justifications for ER Cardinality**

  Description: Point form notes that will provide reasoning for the cardinality choices.

  Due Date: Oct 8, 2025

  Assigned to: Guwinder Khandal

- **Task: Justifications for ER Participation Constraints**

  Description: Point form notes that will provide reasoning for the participation constraints

  Due Date: Oct 8, 2025

  Assigned to: Aaron Herrera

- **Task: Final Draw-up of ER Diagram**

  Description: Final ER diagram will be created and validated by the group

  Due Date: Oct 8, 2025

  Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

Task: The Relational Model

Description: After merging and normalizing, a relational model will be created. Descriptions will be provided on the steps taken to translate the final ER/EER model to a relational model.

Due Date: Oct 9, 2025

Assigned to:

- **Task: Convert ER entities to Relational model**

  Description: Final ER diagram will be created and validated by the group

  Due Date: Oct 8, 2025

  Assigned to: Gurwinder Khandal

- **Task: Convert ER relationships to Relational model**

  Description: Final ER diagram will be created and validated by the group

  Due Date: Oct 8, 2025

  Assigned to: Aaron Herrera

- **Task: Write description regarding the steps to convert to relational model**

  Description: Explain the thought process as to how the ER diagram is to be converted to relational model

  Due Date: Oct 8, 2025

  Assigned to: Jerome De Guzman

- **Task: Final review as a group**

  Description: Ensure relational model reflects the original ER diagram

  Due Date: Oct 9, 2025

  Assigned to: All members will work on this (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: Updated Timeline

Description: The timeline will be updated as a group to make more detailed plans for stages 4-6 and tentative plans for the remaining stages.

Due Date: Oct 10, 2025

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

▼ Reflection #1 → DUE ON OCTOBER 10

Task: Group Reflection/Discussion

Description: 3-5 paragraph reflection written as a group answering the following questions:

- How well is the group communicating? How can communication be improved?
- Has the work been divided evenly among team members?
- Is everyone completing their assigned work on time? If not, what adjustments will be made moving forward to make the project successful?

The group will discuss these topics together to get good idea for each prompt. Notes regarding each prompt will also be taken during the group discussion.

Due Date: Oct 9, 2025

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

- **Task: Writing the paragraph reflection**

  Description: One member will write up the reflection using the notes taken during the discussion

  Due Date: Oct 9, 2025

  Assigned to: Aaron Herrera

Task: Individual Reflection

Description: 3-5 paragraph reflection written individually, answering the following questions:

- Did you complete the tasks you were assigned? Were those tasks completed on time?
- Will you change anything in the way you approach the project moving forward?

- Have you asked for help from your teammates when necessary?
- Is there anything you can do to support your teammates moving forward?

Due Date: Oct 9, 2025

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

▼ Part B: Query & Interface Design

▼ Stage 4) Query check-in (optional submission) → DUE ON OCTOBER 24

Task: Update ER diagram based on previous feedback

Description: An ER diagram to reflect the dataset will be provided. Any feedback will be applied.

Due Date: October 21

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

Task: Brainstorm queries

Description: Each member will come up with 10 queries to potentially implement. For each of the queries give a 1 sentence description.

Due Date: October 22

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

- **Task: Focus on GROUP BY queries**

  Due Date: Oct. 22

  Assigned to: Aaron Herrera

- **Task: Focus on ORDER BY queries**

  Due Date: Oct. 22

  Assigned to: Jerome De Guzman

- **Task: Focus on aggregate function queries**

  Due Date: Oct. 22

Assigned to: Gurwinder Khandal

Task: Create pseudocode for the queries

Description: Write pseudocode for how each query is going to be executed.

Due Date: October 22

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

Task: Review Query & Interface Design

Description: Finalize the list of queries to be implemented as a group

Due Date: Oct. 23

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal

▼ Stage 5) Query Design (5%) → DUE ON NOVEMBER 7

Task: Revise queries based on feedback

Description: Update/revise/create new queries based on the feedback from the optional submission

Due Date: Nov. 5

Assigned to: all members(Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: Write actual code for queries

Description: We will create the code to actually implement our queries. Each member will code a different set of queries (around 10 each).

Due Date: Nov.6

Assigned to:

- **Task: Write actual code for GROUP BY queries**

  Due Date:Nov.6

  Assigned to: Gurwinder Khandal

- **Task: Write actual code for ORDER BY queries**

Due Date: Nov.6

Assigned to: Jerome De Guzman

- **Task: Write actual code for aggregate function queries**

  Due Date:Nov.6

  Assigned to: Aaron Herrera

▼ Stage 6) Interface Design (5%) → DUE ON NOVEMBER 21

Task: Write paragraph about interface description

Description: Write thorough description of interface including how it will look and our plan for implementation.

Due Date: Nov. 18

Assigned to: Aaron Herrera

Task: Review paragraph for interface description

Description: Review the paragraph from the previous task to ensure it accurately describes the interface/plan.

Due Date: Nov. 19

Assigned to: Jerome De Guzman, Gurwinder Khandal

Task: Create diagrams about how interface will look

Description:

Due Date: Nov. 19

Assigned to: Jerome De Guzman and Gurwinder Khandal

- **Task: Create 1 diagram for help menu/instructions**

  Due Date:Nov.19

  Assigned to: Jerome De Guzman

- **Task: Create 1-2 diagrams showing how query results will be presented to user**

Due Date:Nov.19

Assigned to: Gurwinder Khandal

▼ Reflection #2 (3%) → DUE ON NOVEMBER 21

Task: Group Reflection/Discussion

Description: 3-5 paragraph reflection written as a group answering the following questions:

- How well is the group communicating? How can communication be improved?
- Has the work been divided evenly among team members?
- Is everyone completing their assigned work on time? If not, what adjustments will be made moving forward to make the project successful?

The group will discuss these topics together to get good idea for each prompt. Notes regarding each prompt will also be taken during the group discussion.

Due Date: Nov. 20, 2025

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

- **Task: Writing the paragraph reflection**

  Description: One member will write up the reflection using the notes taken during the discussion

  Due Date: Nov. 20, 2025

  Assigned to: Jerome De Guzman

▼ Part C: The Database and its Interface

▼ Database Creation and Population

Task: Discuss feedback from previous stage

Description: The group will reflect on the feedback received. We will make note of any changes that should be made to the query design and interface design.

Due Date: Nov. 26

Assigned to: all members (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: Populate database with data

Description: Using our cleaned data we will use a code-based method to add records into our database to populate it, making sure it is populated properly.

Due Date: Nov.27

Assigned to: Gurwinder Khandal

▼ Implementing an Interface

Task: Create command-line interface

Description: Create front-end interface that allows a person to work on our database, thinking about questions that a top analyst would have in regards to that content in our database, so that we can account for what queries can be used. Securing our database to prevent SQL injections.

Due Date: Nov. 27

Assigned to: Aaron Herrera

Task: Full test of interface and functionality

Description: Each group member will test the implemented interface to ensure that there are no issues (e.g. queries can be ran correctly and display correct output, data population and deletion works, etc.)

Due Date: Nov. 28

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

▼ Project Demonstration (10%) → DECEMBER 1-5

Task: Discuss feedback from previous stage

Description:

Due Date: Nov. 26

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: Schedule a time with instructor to present project

Description: Organize a time as to present project demonstration

Due Date: Before November 20

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: Review feedback from project demonstration

Description: Review feedback, if there are any improvements or adjustments that need to be made to be able to produce a high quality product.

Due Date: Before December 3

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

▼ Final Project Submission (35%) → DUE DECEMBER 5

Task: Write a readme.md file for project .zip file

Description: Write a readme.md file with instructions on how to create and populate the database and run the program.

Due Date: December 3, 2025

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: Organize project files into one .zip file

Description: As a group, we will organize all the files required to create database tables, relationships, populate the tables, and the program written for interacting with the database.

Due Date: December 3, 2025

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

Task: Final review & Submission

Description: Read project grading scale to analyze if everything is completed correctly to achieve the best marks, and read over the submission instructions.

Due Date: December 3, 2025

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

▼ Final Project Report (25%) → DUE DECEMBER 5

Task: Introduction, Discussion of data model draft, and conclusion

Description: Summarize the data model, the logic behind why it was broken down into those tables, issues that arose during designing the model.

Due Date: December 3, 2025

Assigned to: Aaron Herrera

Task: Summary of data and Discussion of the database draft

Description: Summarize the database/data, and create the appendix, consulting members as needed

Due Date: December 3, 2025

Assigned to: Jerome De Guzman

Task: Discussion of the interface draft, list of queries, and appendix

Description: A description of your interface, including a brief description of platform/language used, and screenshots of the interface in action. • A list of interesting queries you can run using the interface. Explain what the queries return, you don’t have to include the SQL code. Explain why these queries would be interesting to an analyst.

Due Date: December 3, 2025

Assigned to: Gurwinder Khandal

Task: Final draft and review

Description: The group will synthesize each part of the report and make any necessary changes for the final draft.

Due Date: December 4, 2025

Assigned to: all members (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

▼ Reflection #3 → DUE DECEMBER 5

Task: Group Reflection/Discussion

Description: 3-5 paragraph reflection written as a group answering the following questions:

- How well did the group communicate?
- Was the work divided evenly among team members?
- Did everyone complete their assigned work on time?
- Were you able to implement the vision that you had for your project earlier in the term? What adjustments were made to the project scope?

The group will discuss these topics together to get good idea for each prompt. Notes regarding each prompt will also be taken during the group discussion.

Due Date: December 4, 2025

Assigned to: All Members will work on this together (Aaron Herrera, Jerome De Guzman, Gurwinder Khandal)

- **Task: Writing the paragraph reflection**

  Description: One member will write up the reflection using the notes taken during the discussion

Due Date: Dec. 4, 2025

Assigned to: Gurwinder Khandal

Task: Individual Reflection

Description: 3-5 paragraph reflection written individually, answering the following questions:

- Did you complete the tasks you were assigned?
- Were those tasks completed on time?
- Did you provide support to your teammates? How and when?
- Did you ask for help from your teammates when necessary?
- The next time you work on a team, is there anything you will do differently?

Due Date: December 4, 2025

Assigned to: Aaron Herrera, Jerome De Guzman, Gurwinder Khandal