Jacob Seraspi
Joshua Penner
Sahil Sharma

Stage 1: Dataset Selection and Group Timeline

Dataset: Please see the link below for the dataset.
https://catalog.data.gov/dataset/subcontracting-directory-for-small-businesses

The dataset we are utilizing is the General Services Administration's subcontracting directory for small businesses. It encompasses detailed information on government contracts, focusing on various attributes associated with each awarded contract. Key features include the Procurement Instrument Identifier (PIID), estimated ultimate completion dates, and specifics about contracting and funding agencies. It captures the nature of the work through NAICS codes and descriptions, as well as product or service codes.

The dataset contains numerous records, each representing a unique government contract. Additional attributes include the period of performance, principal place of performance (city, state, and zip code). Contractor details are also present, including the contractor's name, address, and contact information. Additionally, the dataset highlights unique entity IDs and includes information about business size determinations and subcontract plans.

Overall, this dataset serves as a valuable resource for analyzing government contracting practices. It provides insights into the types of services procured, the agencies involved, and geographical trends. Its structured and detailed format makes it ideal for further analysis, such as identifying patterns in contracting behavior and evaluating the impact of government spending on various industries.

This dataset needs cleaning as it contains redundant data within many columns. For instance, under the attribute “Funding Agency Name,” there are repeated names of the agencies that need to be removed by using the database management principles. Moreover, to remove the redundant data and make the dataset more standardized, we are planning to use Microsoft Excel as a tool to clean data.

Jacob Seraspi
Joshua Penner
Sahil Sharma

Project Timeline

Task / Subtask	Due Date	Assignee
Dataset Choice	Sept 20	All
Project Summary		Jacob Sahil
Project Timeline		Josh
ER Diagram (opt)	Sept 27	All
Reminder of chosen data		Josh
ER Diagram		Jacob
Justification of ER Diagram decisions		Sahil
Database Design	Oct 11	All
Data summary		Sahil
ER / EER Diagram		Jacob
Final relational model & justification		Josh
Reflection #1	Oct 11	All
Query Ideas (opt)	Oct 24	All
List of queries to implement		All
Queries	Nov 8	All
Updates to queries		Josh
Query justifications		Sahil / Jacob
Interface Design	Nov 22	All
Description of Interface and implementation plan		Sahil / Josh
Diagrams of Interface		Jacob
Reflection #2	Nov 22	All
Project Demonstration	Nov 29 - Dec 6	All
Schedule Instructor meeting		TBD
Final Submission / Part C	Dec 6	All

Interface Creation		TBD
UI / DB Integration		TBD
Readme file		TBD
Code to populate DB		TBD
Final Project Report	Dec 6	All
Reflection #3	Dec 6	All

Subtask Descriptions

Note: For main task descriptions, see COMP 3380 Project Instructions document

- Dataset Choice
  - Project Summary
    - 3-5 paragraph summary of the dataset chosen and why it was chosen. Includes list of attributes and data readiness.
  - Project Timeline
    - Outline of tasks to be completed for the project. Detailed schedule only required for stages 1-3.
- ER Diagram (opt)
  - Reminder of chosen data
    - Summary of data, updated with stage 1 feedback
  - ER Diagram
    - ER Diagram of data, more detail given allows for more feedback.
  - Justification of ER Diagram decisions
    - Justify relationships, participation, and cardinality
- Database Design
  - Data summary
    - Update of summary from stage one, reflecting feedback received
  - ER / EER Diagram
    - ER / EER Diagram with text justifying constraints
  - Final relational model & justification
    - Include final relational model
    - Description of steps taken to merge and normalize final relational model
- Query Ideas (opt)
  - List of queries to implement
- Query Design
  - Updates to queries
    - Make changes to the queries based on feedback from Stage 4
  - Query justifications
    - Explanations of why the results would be interesting to an analyst
- Interface Design

Jacob Seraspi
Joshua Penner
Sahil Sharma

- Description of interface and implementation plan
  - Describe the look of the interface and the plan to implement it, including language, CLI vs GUI, etc.
- Diagrams of interface
  - Diagram showing how the interface will look, including help menu and how the results will be displayed to the user.

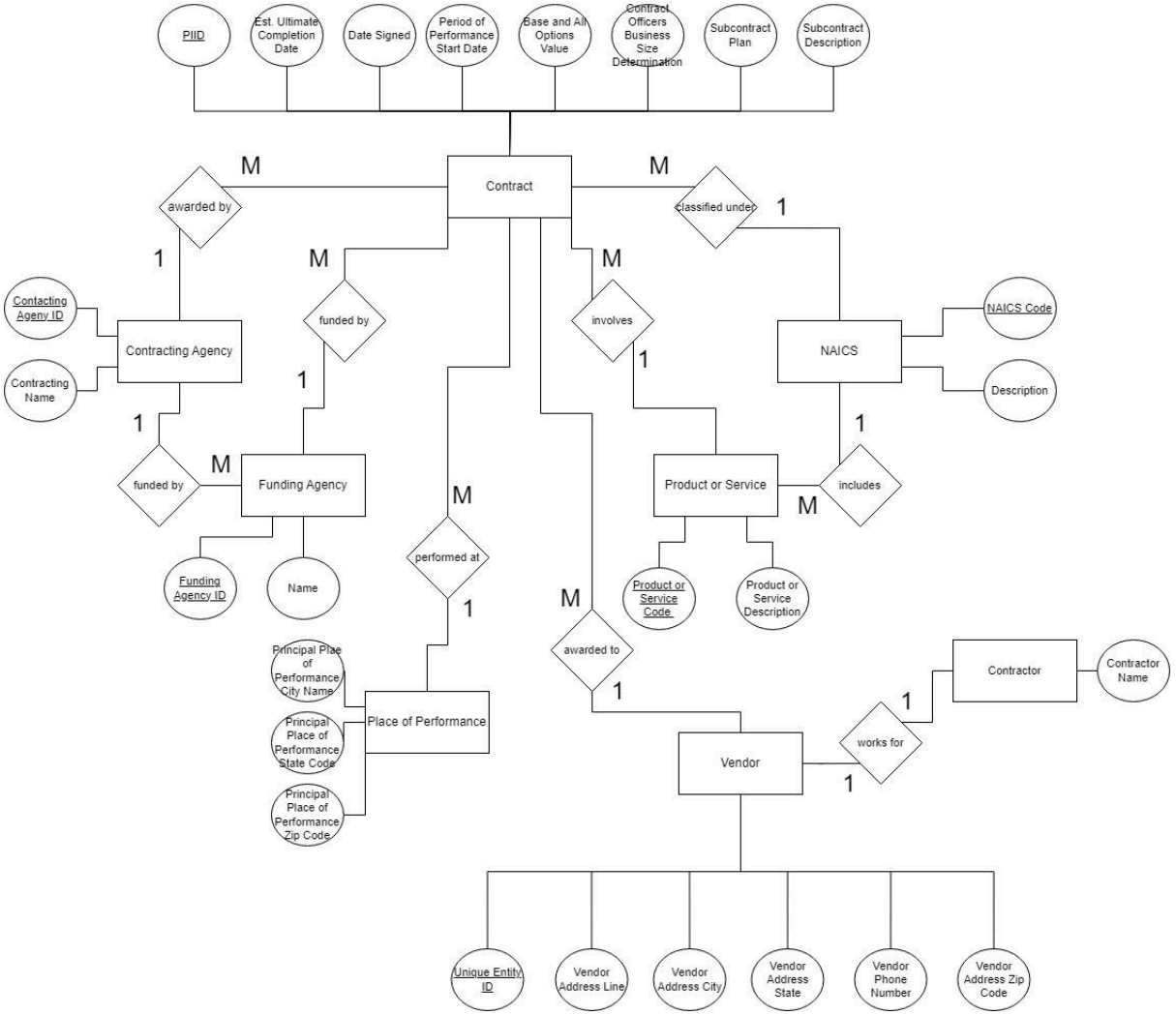
Jacob Seraspi
Joshua Penner
Sahil Sharma

Stage 2: ER Diagram

Summary of data:

The General Services Administration's subcontracting directory for small businesses provides comprehensive data on government contracts, including contract details, agency information, contractor details, and business size determinations. This dataset is valuable for analyzing government contracting practices and identifying patterns in contracting behavior. However, the dataset contains redundant data, particularly in the "Funding Agency Name" column, which needs to be cleaned using database management principles. Microsoft Excel will be used as a tool to remove redundant data and standardize the dataset.
No feedback has been provided from stage 1, so no updates based on said feedback can be made.

ER Diagram:



Jacob Seraspi
Joshua Penner
Sahil Sharma

Justification of ER Diagram decisions:

Contracting Agency and Funding Agency
The funded by relationship is one to many because each contracting agency can be funded by many funding agencies and each funding agency will have one contracting agency. Contractor agencies have partial participation because they could exist in the database without an assigned funding agency and similarly funding agency have a partial participation as it could exist without having a contractor agency in the database.

Contracting Agency and Contract
The awarded by relationship is one to many because each contracting agency will have many contracts, and each contract will have one contracting agency. Contracting agency have total participation because a contracting agency must have contracts to exist in database and contracts have partial participation because it could exist in the database without the assigned contracting agency.

Funding Agency and Contract
Funded by relation for funding agency and contract entities is one to many because each funding agency can provide funds to multiple contracts and each contract can be funded by one funding agency. Funding agency have total participation because funding agency must have contracts to fund and contracts have partial participation because they could exist in the database without an assigned funding agency.

Contract and Place of Performance
The performed at the relationship is one to many because each place of performance will have multiple contracts whereas each contract will have only one place of performance. Place of performance has partial participation because it could exist in a database without an assigned contract. The contract has partial participation because it could exist in a database without place of performance.

Contract and Vendor
The awarded relationship for contract and vendor entities has one to many because each vendor can have multiple contracts, and each contract will have one vendor. Vendors have total participation because a database vendor could exist without the assigned contract and the contract has partial participation because it could exist in the database without the assigned vendor.

Contract and Product or Service
The involved relationship is one to many because each contract will have one product or service, and each product or service will have many contracts. Products or Service have total participation because a database product or service could exist

without an assigned contract and the contract has partial participation because it could exist in a database without assigned product or service.

Contract and NAICS
The classified under relationship is one to many because each contract will have one NAICS code and each NAICS code can have multiple contracts. Contracts have partial participation because they could exist in databases without assigned NAICS code and NAICS have total participation because we don't want to store NAICS in the database until it has a contract.

NAICS and Product or Service
The include relationship is one to many because each NAICS code has many products or services whereas each product or service has one NAICS code. NAICS have partial participation because it could exist in a database without assigned product or service and product or service have partial participation because it could exist in a database without assigned NAICS.

Vendor and Contractor
The works relationship is one to one because each vendor has one contractor, and each contractor has one vendor at a time. The contractor has total participation because the contractor can exist in the database only if they have assigned a vendor and vendor has partial participation because it could exist in the database without the need of the contractor.

Stage 3: Database Design

Data Summary:

Dataset: Please see the link below for the dataset.

https://catalog.data.gov/dataset/subcontracting-directory-for-small-businesses

We are utilizing a subcontracting directory specifically for small businesses. This dataset features a list of prime contractors who must develop plans and goals for subcontracting with small business firms. It consists of several columns, some of which contain unique IDs for specific contractors. The information in this dataset is highly valuable, and we believe that developing a database from it will assist

contractors in building connections with clients and discussing the goals and requirements necessary for achieving profitable business.

The attributes for this dataset are as follows:

1) PIID

2) Est. Ultimate Completion Date

3) Contracting Agency Name

4) Contracting Agency ID

5) Funding Agency ID

6) Funding Agency Name

7) Date Signed

8) NAICS Code

9) NAICS Description

10) Product or Service Code

11) Product or Service Description

12) Period of Performance Start Date

13) Principal Place of Performance City Name

14) Principal Place of Performance State Code

15) Place of Performance Zip Code

16) Contractor Name

17) Vendor Address Line 1

18) Vendor Address City

19) Vendor Address State

20) Vendor Phone Number

21) Vendor Address Zip Code

22) Unique Entity ID

23) Contracting Officers Business Size Determination

Jacob Seraspi
Joshua Penner
Sahil Sharma

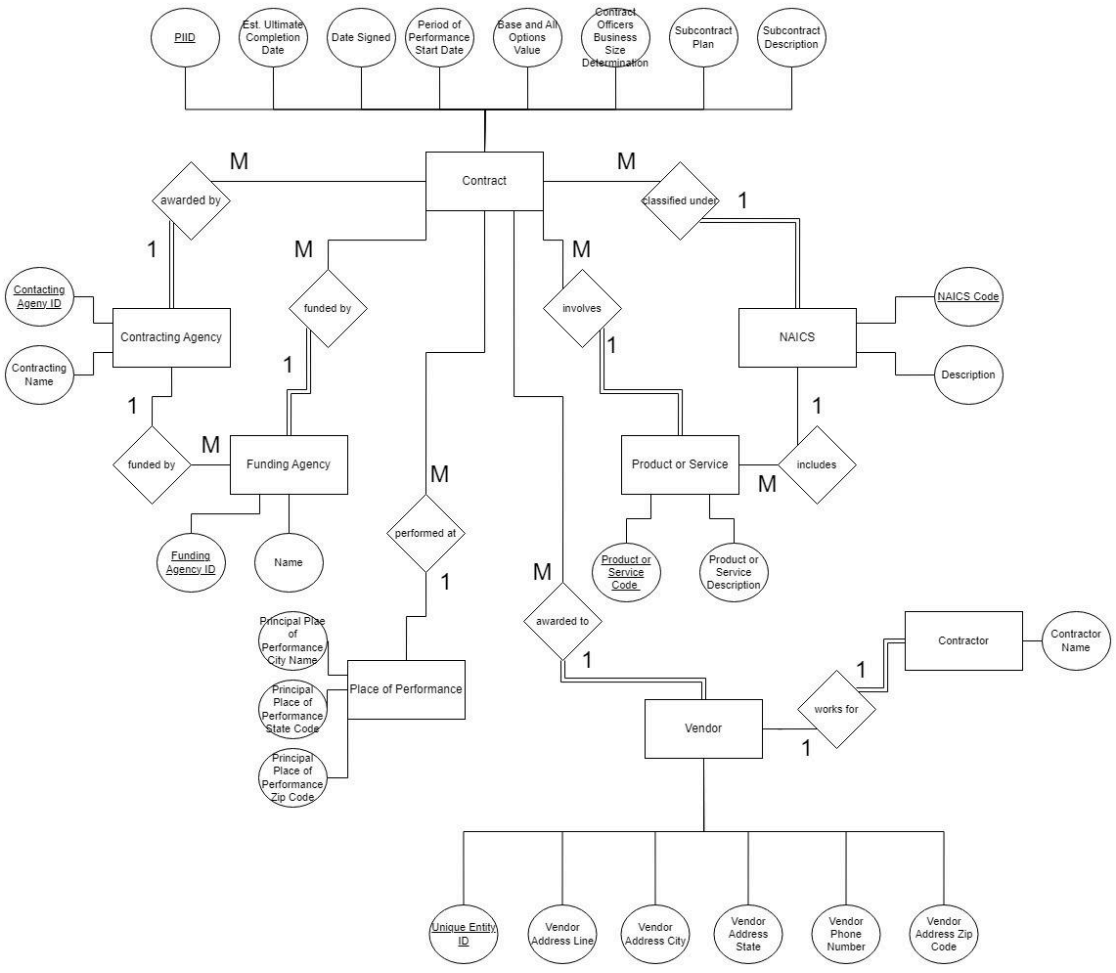
24) Subcontract Plan

25) Subcontract Plan Description

26) Base and All Options Value (Total Contract Value)

Overall, this dataset is a valuable resource for analyzing government contracting practices. It offers insights into the types of services procured, the agencies involved, and geographical trends. Its structured and detailed format makes it well-suited for further analysis, including identifying patterns in contracting behavior and assessing the impact of government spending on different industries.

ER Diagram:



Jacob Seraspi
Joshua Penner
Sahil Sharma

Justification of ER Diagram decisions:

Contracting Agency and Funding Agency
The funded by relationship is one to many because each contracting agency can be funded by many funding agencies and each funding agency will have one contracting agency. Contractor agencies have partial participation because they could exist in the database without an assigned funding agency and similarly funding agency have a partial participation as it could exist without having a contractor agency in the database.

Contracting Agency and Contract
The awarded by relationship is one to many because each contracting agency will have many contracts, and each contract will have one contracting agency. Contracting agency have total participation because a contracting agency must have contracts to exist in the database and contracts have partial participation because it could exist in the database without the assigned contracting agency.

Funding Agency and Contract
Funded by relation for funding agency and contract entities is one to many because each funding agency can provide funds to multiple contracts and each contract can be funded by one funding agency. Funding agency have total participation because funding agency must have contracts to fund and contracts have partial participation because they could exist in the database without an assigned funding agency.

Contract and Place of Performance
The performed at the relationship is one to many because each place of performance will have multiple contracts whereas each contract will have only one place of performance. Place of performance has partial participation because it could exist in a database without an assigned contract. The contract has partial participation because it could exist in a database without place of performance.

Contract and Vendor
The awarded relationship for contract and vendor entities has one to many because each vendor can have multiple contracts, and each contract will have one vendor. Vendors have total participation because a database vendor could exist without the assigned contract and the contract has partial participation because it could exist in the database without the assigned vendor.

Contract and Product or Service
The involved relationship is one to many because each contract will have one product or service, and each product or service will have many contracts. Products or Service have total participation because a database product or service could exist

without an assigned contract and the contract has partial participation because it could exist in a database without assigned product or service.

Contract and NAICS
The classified under relationship is one to many because each contract will have one NAICS code and each NAICS code can have multiple contracts. Contracts have partial participation because they could exist in databases without assigned NAICS code and NAICS have total participation because we don't want to store NAICS in the database until it has a contract.

NAICS and Product or Service
The include relationship is one to many because each NAICS code has many products or services whereas each product or service has one NAICS code. NAICS have partial participation because it could exist in a database without assigned product or service and product or service have partial participation because it could exist in a database without assigned NAICS.

Vendor and Contractor
The works relationship is one to one because each vendor has one contractor, and each contractor has one vendor at a time. The contractor has total participation because the contractor can exist in the database only if they have assigned a vendor and vendor has partial participation because it could exist in the database without the need of the contractor.

Functional Dependencies

Contract
- PIID
    - Subcontract Plan Description
    - Subcontract Plan
    - Product or Service Code
    - Contracting Officers Business Size Determination
- Subcontract Plan
    - Subcontract Plan Description

Contracting Agency
- Contracting Agency ID
    - Contracting Agency Name

NAICS
- NAICS Code
    - NAICS Description

Product or Service
- Product or Service Code
    - Product or Service Description

Place of Performance

Jacob Seraspi
Joshua Penner
Sahil Sharma

- Place of Performance Zip Code
    - Principal Place of Performance City Name
    - Principal Place of Performance State Code

Vendor

- Unique Entity ID
    - Vendor Address Line 1
- Vendor Address Line 1
    - Vendor Address Zip Code
    - Vendor Address City
    - Vendor Address State
- Vendor Phone Number
    - Unique Entity ID (Idk why but some Entity IDs have multiple phone numbers so vendor phone number can be used to determine unique entity id)

Normalization

1NF:

- PIID → Subcontract Plan, Subcontract Plan Description, Contracting Agency ID, Product or Service Code, NAICS Code, Place of Performance Zip Code
- Subcontract Plan → Subcontract Plan Description
- Contracting Agency ID → Contracting Agency Name
- Product or Service Code → Product or Service Description
- NAICS Code → NAICS Description
- Place of Performance Zip Code → Principal Place of Performance City Name, Principal Place of Performance State Code
- Unique Entity ID → Vendor Address Line 1, Vendor Address City, Vendor Address State, Vendor Address Zip Code
- Vendor Phone Number → Unique Entity ID

2NF:

- Contract(PIID, Subcontract Plan, Contracting Agency ID, Product or Service Code, NAICS Code, Place of Performance Zip Code)
- Contracting Agency(Contracting Agency ID, Contracting Agency Name)
- Product or Service(Product or Service Code, Product or Service Description)
- NAICS(NAICS Code, NAICS Description)
- Place of Performance(Place of Performance Zip Code, Principal Place of Performance City Name, Principal Place of Performance State Code)
- Vendor(Unique Entity ID, Vendor Address, Line 1, Vendor Address City, Vendor Address State, Vendor Address Zip Code, Vendor Phone Number)

3NF:

- There are no transitive dependencies, so there's no need for change in the tables.

BCNF:

- The primary keys are the same and all the determinants are superkeys, so there are no required changes for the tables.

Jacob Seraspi
Joshua Penner
Sahil Sharma

Summary of Normalization:

1NF

- We made sure each attribute is atomic and made sure that there are no multi-valued or repeating groups.

2NF

- We eliminated partial dependencies by splitting the attributes into separate tables such as Contracting Agency, Product or Service, and Place of Performance.

3NF

- No changes necessary because there's no transitive dependencies

BCNF

- No changes necessary because all the determinants are superkeys.

Updated Timeline

Task / Subtask	Due Date	Tasks status	Assignee
Dataset Choice	Sept 20	Completed	All
Project Summary		Completed	Jacob Sahil
Project Timeline		Completed	Josh
ER Diagram (opt)	Sept 27	Completed	All
Reminder of chosen data		Completed	Josh
ER Diagram		Completed	Jacob

Jacob Seraspi
Joshua Penner
Sahil Sharma

Justification of ER Diagram decisions		Completed	Sahil
Database Design	Oct 11	Completed	All
Data summary		Completed	Sahil
ER / EER Diagram		Completed	Jacob
Final relational model & justification		Completed	Josh
Reflection #1	Oct 11	Completed	All
Query Ideas (opt)	Oct 25	Pending	All
List of queries to implement	Expected completion – November 5	Pending	All
Queries	Nov 8	Pending	All
Interface Design	Nov 22	Pending	All
Reflection #2	Nov 22	Pending	All
Project Demonstration	Nov 29 - Dec 6	Pending	All

Jacob Seraspi
Joshua Penner
Sahil Sharma

Final Submission	Dec 6	Pending	All
Final Project Report	Dec 6	Pending	All
Reflection #3	Dec 6	Pending	All

Jacob Seraspi
Joshua Penner
Sahil Sharma

The NBA Database is an expansive and detailed dataset that encompasses professional basketball from the league's inception in 1946-47 to the present, covering over 65,000 games, more than 4,800 players, and over 13 million rows of play-by-play events. This data serves as a resource for analyzing a wide range of basketball elements, from player performance and game outcomes to historical trends. Stored as an SQLite file, the database captures key entities such as **Players**, **Draft Picks**, **Draft Combine Stats**, **Teams**, **Games**, **Officials**, **Seasons**, **Events**, and **Quarters**—each providing critical insights into the NBA's multifaceted data landscape.

Each **Player** entity includes attributes such as player ID and name, which allow users to perform player-specific analyses and link them to game outcomes and draft details. **Draft Picks** represent the selection process for players, with attributes like round number, overall pick, and draft type, enabling users to track how a player's draft position correlates with their career progress. The **Draft Combine Stats** table includes measurements and physical performance metrics, such as height, weight, wingspan, vertical jump, and bench press performance, providing valuable context for player development and fitness standards at the start of their careers.

Teams are defined by a team ID, name, and abbreviation, creating a clear structure for tracking team affiliations across seasons. Each **Game** entity links players and teams, recording essential details like game ID, date, and final scores, including win/loss information for home teams. **Officials** track the referees who participate in games, with each official characterized by an ID, name, jersey number, and other identifying details. This linkage between officials and games offers insights into officiating trends and their potential impact on game dynamics and outcomes.

The **Seasons** entity organizes games by year and season type, enabling an overview of trends across different NBA periods, from regular seasons to playoffs. **Events** are logged for each game, with attributes such as event number, description, message type, and timing (WC Time and PC Time), providing granular details on each play's outcome and sequence. Finally, **Quarters** divide games into four parts, allowing analysts to dissect game flow, scoring patterns, and strategy adjustments made by teams over time.

Together, these entities and attributes form a robust relational structure, ideal for exploring correlations between player stats, team dynamics, game results, and historical league trends. The organized structure of this dataset, along with its comprehensive functional dependencies, supports high-integrity analysis, making it a critical tool for understanding the evolution of the NBA, predicting game outcomes, and modeling player performance.

Players

- **Player ID**: Unique identifier for each player.
- **Player Name**: Full name of the player.

Draft Picks

- **Round number**: The round in which the player was drafted.
- **Round pick**: The specific pick number within the round.

- **Overall pick**: The player's overall pick number in the draft.
- **Draft type**: Type of draft, indicating if it was a regular or supplemental draft.

Draft Combine Stats

- **Season**: The NBA season during which the combine stats were recorded.
- **Height (in)**: Player's height in inches.
- **Weight (lbs)**: Player's weight in pounds.
- **Wingspan (in)**: The player's wingspan in inches.
- **Standing Reach (in)**: Measurement of the player's reach from the ground to the top of their hand when standing.
- **Body Fat %**: Player's body fat percentage.
- **Standing Vertical**: Measurement of the player's vertical jump from a standing position.
- **Max Vertical**: The player's maximum vertical jump height.
- **Bench Press**: The number of repetitions of a set weight the player can lift.

Teams

- **Team ID**: Unique identifier for each team.
- **Team name**: Full name of the team.
- **Team Abbreviation**: Shortened abbreviation used for the team.

Games

- **Game ID**: Unique identifier for each game.
- **Game Date**: Date when the game was played.
- **W/L Home**: Win or loss result for the home team.
- **Points Home**: Total points scored by the home team.
- **Points Away**: Total points scored by the away team.

Officials

- **Official ID**: Unique identifier for each official.
- **First Name**: First name of the official.
- **Last Name**: Last name of the official.
- **Jersey Number**: Jersey number worn by the official.

Seasons

- **Season ID**: Unique identifier for each NBA season.
- **Season Type**: Type of season, indicating regular season, playoffs, or other.
- **Year**: Year(s) that the season took place.

Events

- **Event Number**: Unique identifier for each event within a game.

- **Description**: Text description of the event, detailing the play or action.
- **Event Message Type**: Type or category of event message, such as a shot, foul, or substitution.
- **WC Time**: Game clock time recorded in Wall Clock format.
- **PC Time**: Time recorded in Period Clock format, showing the remaining time in the quarter.

Quarters

- **Quarter**: Division of a game representing each of the four periods in regulation time.

Jacob Seraspi
Joshua Penner
Sahil Sharma

Updated Timeline

Task / Subtask	Due Date	Tasks status	Assignee
Dataset Choice	Sept 20	Completed	All
Project Summary		Completed	Jacob Sahil
Project Timeline		Completed	Josh
ER Diagram (opt)	Sept 27	Completed	All
Reminder of chosen data		Completed	Josh
ER Diagram		Completed	Jacob
Justification of ER Diagram decisions		Completed	Sahil
Database Design	Oct 11	Completed	All
Data summary		Completed	Sahil
ER / EER Diagram		Completed	Jacob
Final relational model & justification		Completed	Josh

Jacob Seraspi
Joshua Penner
Sahil Sharma

Reflection #1	Oct 11	Completed	All
Query Ideas (opt)	Oct 25	Pending	All
List of queries to implement	Expected completion – November 5	Pending	All
Queries	Nov 8	Pending	All
Interface Design	Nov 22	Pending	All
Reflection #2	Nov 22	Pending	All
Project Demonstration	Nov 29 - Dec 6	Pending	All
Final Submission	Dec 6	Pending	All
Final Project Report	Dec 6	Pending	All
Reflection #3	Dec 6	Pending	All

Jacob Seraspi
Joshua Penner
Sahil Sharma