**Big Data Project**

**What is JEOPARDY!**



**Megan Ball, Matt Farrow, & Jake Harrison**

What is **Abstract**:

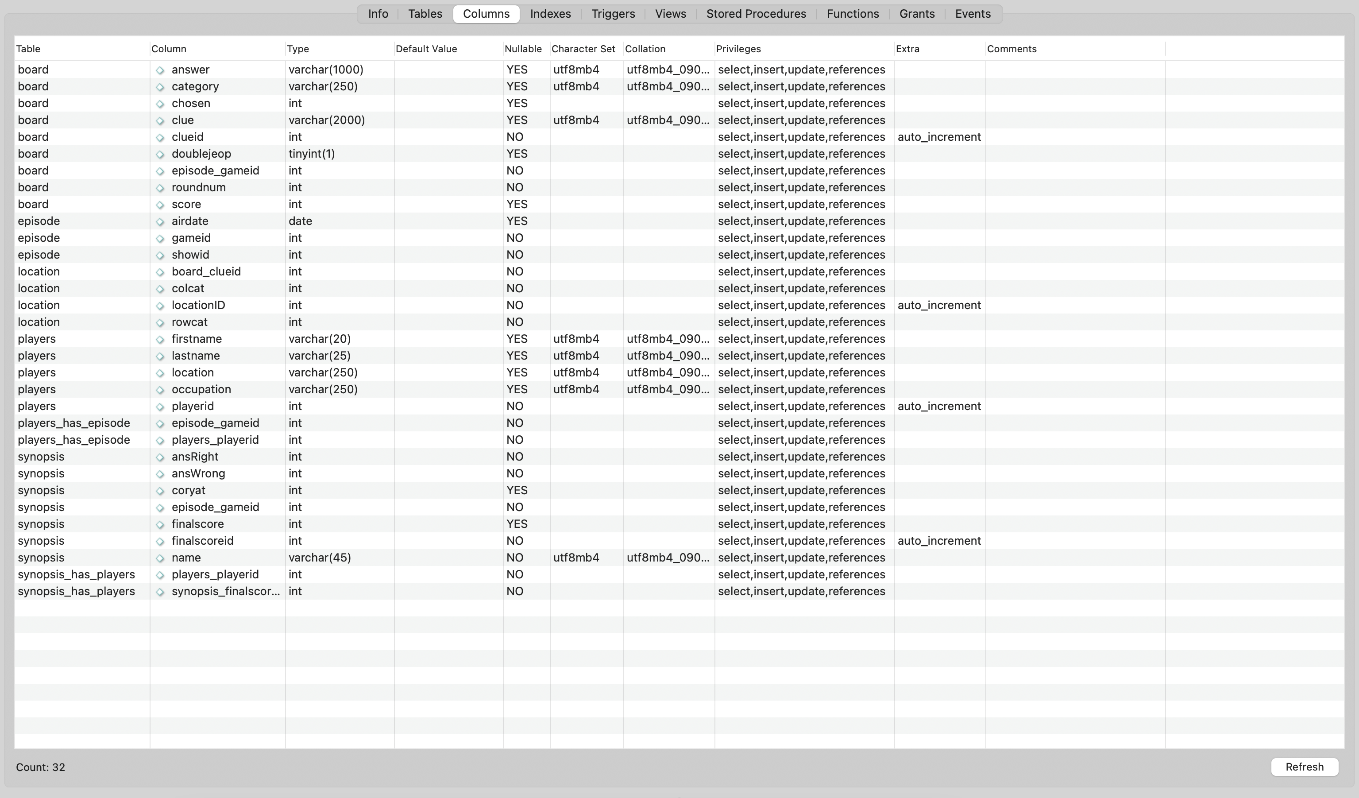
For this project, the team’s primary goal is to create a database using JEOPARDY! data and visualize insights. For the team to proceed, the team had to extract the data from the website, [J! Archive (j-archive.com)](https://j-archive.com/) using R script. Using MySQL Workbench, the schema was created and the database was populated. Once complete, the team will be able to connect the database with an RShiny application.

What is **Key Goals**:

The key goals for this project were obtaining the data from the source, creating the database schema, manipulating the data to populate the database, populating the database, gathering insights from games and notable players, and utilizing the database to create an RShiny app (TBD).

What is **Data**:

The JEOPARDY! data contained an abundant amount of information regarding the categories, clues, answers, scores, players, and much more. Figure 1 shows all the columns used in our database.

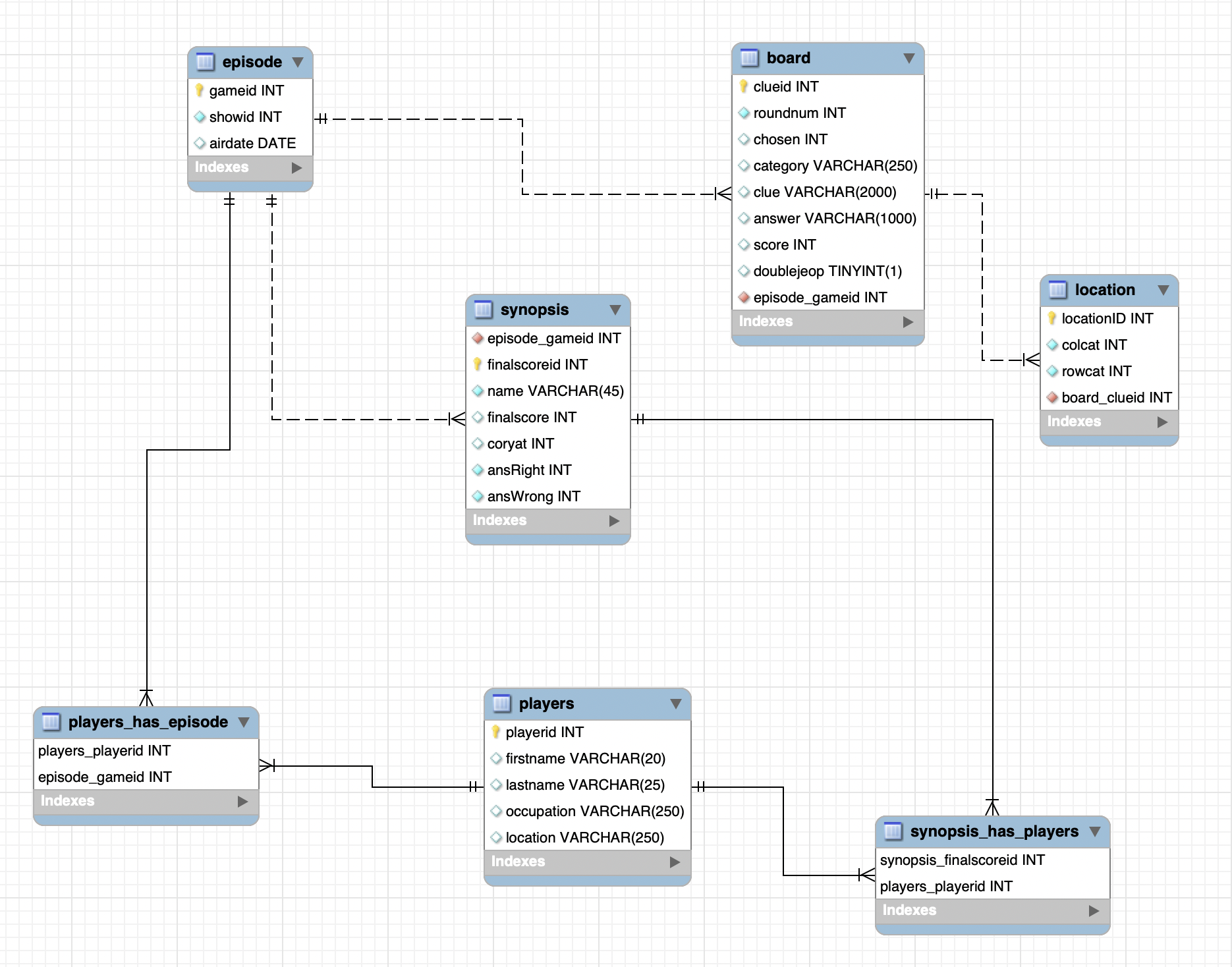


*Figure 1. Columns of the entire JEOPARDY! database with their data types.*

The `whatr` package was used for the majority of the data collection for this project. Using the package's included functions, we were able to crawl the J! Archive and extract the air date, board details, daily double information, final scores, player information, and game synopsis for 6,948 games spanning from 1983 until 2021. Once the data had been crawled, a function with a `for` loop was used to iterate over each game and bind the rows of each piece of information into a single data set.

What is **Database Structure**:

Once the team understood which variables were pertinent to the team’s overall goal, the team was able to create a normalized schema which is represented by the enhanced entity-relationship diagram in Figure 2. There are two many-to-many relationships within the schema between the tables *players* and *episodes*, as well as the *synopsis* and *players*. Therefore, the team had to create specialized tables that could represent these many-to-many relationships: *players\_has\_episode* and *synopsis\_has\_players*. The other relationships are one-to-many/many-to-one.



*Figure 2. The Enhanced Entity-Relationship Diagram of the JEOPARDY! database.*

Based on the schema in the enhanced entity-relationship diagram, the team determined the database was normalized . The data was manipulated in R to satisfy the schema that had been design in MySQL Workbench, and was then uploaded into the database.

What is **Future Work**:

The future goals of the team are providing data insights from games and notable players and creating the code for the RShiny application.

To find the work that has already been completed, please refer to the github link:

[mattfarrow1/7330-term-project: Final project for DS 7330 File Organization and Database Management (github.com)](https://github.com/mattfarrow1/7330-term-project)