Analyzing *Jeopardy!* Data

M. Ball, M. Farrow, J. Harrison

***Abstract*—For this project, the team's primary goal was to create a database using Jeopardy! data and visualize insights. For the team to proceed, the team used R to crawl the website, J! Archive [1] and extract the data. Using MySQL Workbench, a schema was created and the database was populated. Once complete, the team connected to the database with an RShiny application.**

***Index Terms*—database, relational database, text analysis, text mining**

I. Introduction

JEOPARDY! is a long running quiz-style television game show where contestants are presented clues in the form of answers from a series of categories and must phrase their responses in the form of a question. Each game is composed of three rounds: Jeopardy, Double Jeopardy, and Final Jeopardy. The first two rounds are made up of six categories with five answers of increasing difficulty and monetary value. If a contestant provides the correct question, they receive the amount that question is worth; otherwise, they lose that amount. Although the original version of *Jeopardy!* premiered in 1964, the show's revival in 1984 with host Alex Trebek led to a run of more than 8,000 games over 37 years [2].

Using data from 6,775 of those games, we wanted to analyze the data provided on J! Archive to examine what has made this show such a game show staple. We examined trends across game categories, player appearances, and Daily Doubles.

II. Existing Research

Jeopardy! is a game show that lends itself to trends and statistical studies, and trivia's broad appeal gives the game accessibility to players and viewers alike. Sites such as The Jeopardy! Fan [3] and the Jeopardy! History Wiki [4] have provided facts and statistics from the game, with data often coming from J! Archive. Even sites like FiveThirtyEight have turned their analytical prowess on the game, analyzing two of the show's most notable players, Ken Jennings and James Holzhauer [5].

The J!-Archive has enabled all of this work, but one of the site's limitations is its ability to interact with the data. Our intention with this project was to take the data from J!-Archive and create a database that could be used by researchers to analyze the data for themselves.

III. Data

Our approach to this project involved crawling the J!-Archive website and balancing the wealth of available data with the goals of this project. Due to the connected nature of the data, we made the decision to build our project in a relational database using MySQL. Initial data collection and subsequent analysis took place using the R programming language. All source data files and analysis queries are located on GitHub [6].

A. Data Collection

The `whatr` [7] 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 almost 7,000 games spanning more than three decades. 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.

B. Data Models

Once we understood which variables were pertinent to the project's overall goal, we were able to create a normalized schema which is represented by the enhanced entity-relationship diagram (Figure 1). There are three many-to-many relationships within the schema: one between players and episodes, another between Daily Doubles and their scores, and finally between the game synopsis and players. Therefore, the team had to create specialized tables that could represent these many-to-many relationships: *players_has_episode*, *doubles_has_scores,* and *synopsis_has_players*. The other relationships are one-to-many/many-to-one.

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 designed in MySQL Workbench, and was then uploaded into the database.

This paper was submitted as the final term project for SMU's MSDS 7330: File Organization and Database Management course under the supervision of Dr. Sohail Rafiqi.

Megan Ball is with Southern Methodist University, Dallas, TX 75025 USA (e-mail: ballm@smu.edu).

Matt Farrow is with Southern Methodist University, Dallas, TX 75025 USA. (e-mail: mfarrow@smu.edu).

Jake Harrison is with Southern Methodist University, Dallas, TX 75025 USA (e-mail: harrisonjp@mail.smu.edu)

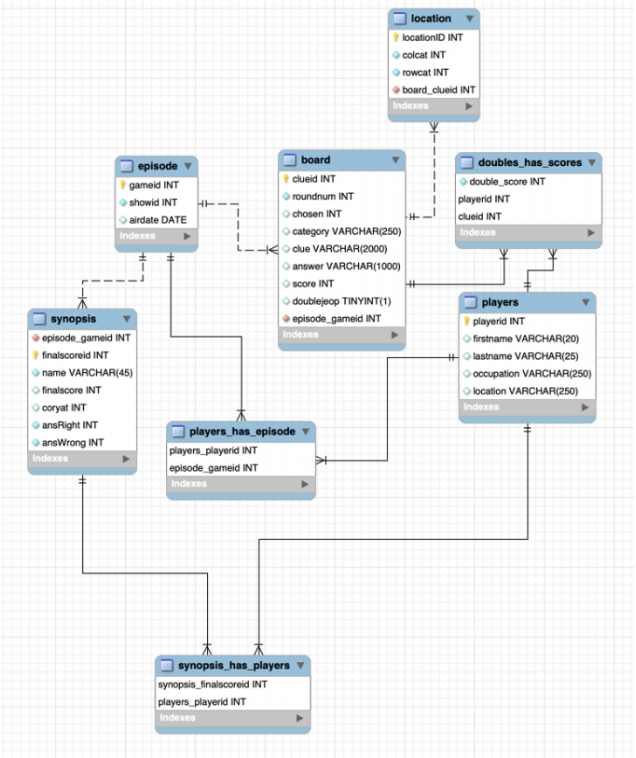

Figure 1: EER Diagram

IV. Analysis

Once the database was complete, the next steps were to perform exploratory data analysis and note key insights from both the history of the game and from the notable players.

A. Top Players

Before discussing any other key items, it is important to note the top players in order to understand some of the key trends we discovered within the data. James Holzhauer, a professional sports gambler from Las Vegas, Nevada, holds the record for the highest single-game score of $131,127. There is a significant delta between him and the next top single-game scorer, Ken Jennings, who maxed out at only $75,000 in a single game. However, Ken Jennings is typically a more well-known *Jeopardy!* contestant as he holds the record for the most consecutive appearances on the show at 75.

During *Jeopardy!*'s 20th anniversary season which began in the fall of 2003, the rules changed and set no limit on a returning champion's number of consecutive appearances [8]. Prior to this rule change, a returning champion could only appear on a maximum of five consecutive episodes. Therefore our notable player statistics skew towards all contestants that competed during or after the 20th anniversary season (Table I). For the purposes of this paper, the 'top 10 players' mentioned later on will refer to this list of players.

TABLE I
NOTABLE PLAYER STATISTICS*

Player	Highest Score	Cumulative Correct Answers	Cumulative Incorrect Answers	Total Number of Games
James Holzhauer	$131,127	1154	35	33
Ken Jennings	$75,000	2643	240	75
Jason Zuffranieri	$58,400	565	34	29
Julia Collins	$35,000	504	42	21
David Madden	$34,200	470	36	20
Matt Jackson	$51,000	389	14	14
Austin Rogers	$69,000	322	42	13
Arthur Chu	$58,200	309	42	12
Seth Wilson	$31,200	307	30	13
Jason Keller	$36,900	246	24	10

*all statistics exclude any tournament or special game series

B. Categories

One of the first questions that we asked was, "Are there common categories that appear in multiple games?" Figure 2 shows the 10 most common categories that appear in the database and the number of games that they appear in. Unsurprisingly, the most popular categories tend to be broad, offering the game designers significant flexibility to design clues. However, it also indicates that there are certainly recurring themes within the game, and it would be wise to study key items within these subjects in preparation for the game.

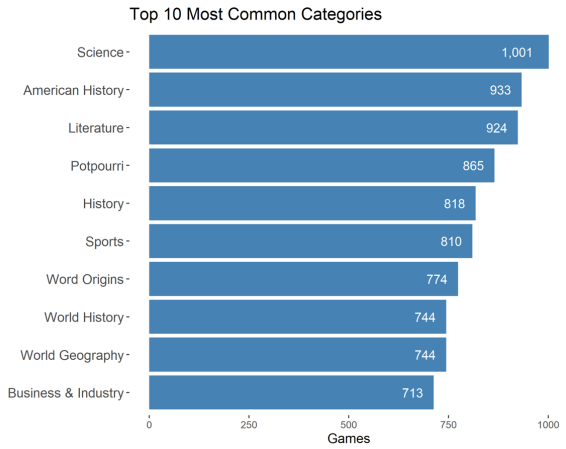

Figure 2: Most common categories

C. Daily Doubles

A Daily Double in *Jeopardy!* is a clue that allows the player who selected the clue the opportunity to wager an amount of money, from $5 up to their current score, before seeing the answer [9]. If the contestant gets the question correct, their wager is added to their current score. If the contestant gets the question wrong, the wager is subtracted from their current score.

In each game, a single Daily Double appears in the first round, and two Daily Doubles appear in the second round. Historically, Daily Doubles tended to appear as contestants worked their way down a category, but contestants like James

Holzhauer have bucked that approach and built their appearance strategies around actively seeking out Daily Double clues [10].

An analysis of Daily Double locations revealed a strong placement trend, particularly along the fourth row, that more strategic players may seek out in an effort to gain a financial advantage over their opponents. (Figure 3)

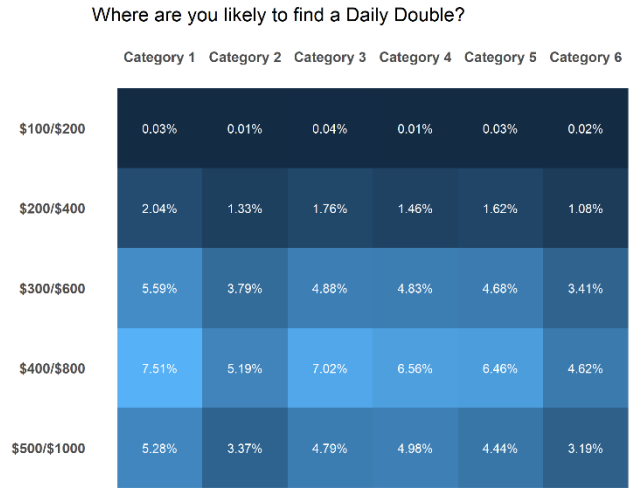

Figure 3: Plot of Daily Double locations

Among the top ten players, the location in row 4 and column 3 was chosen the most at 39 times, followed closely by the location in row 4 and column 1 at 36 times (both $400 single jeopardy or $800 double jeopardy clues). This aligns with the two clue locations with the highest frequency of having a daily double clue at 7.02% and 7.51%, respectively.

Figure 4 shows a word cloud made from Daily Double clues, the most common being "name", "one", or "named." Looking at the words, it seems that players who come across a Daily Double are more likely to be asked to identify something or someone, as opposed to pinpointing a date or time in history.

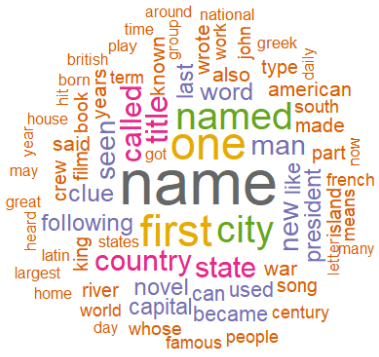

Figure 4: Word frequency in Daily Double clues

A word cloud of Daily Double answers seems to bear that hypothesis out, as seen in Figure 5. A safe estimate would be that British or American historical figures are a common theme, as we see "John," "William", "George", and "Thomas" appear very often.

Figure 5: Word frequency in Daily Double answers

When examining who got the most daily double clues correct, it is not surprising to find that the players align almost perfectly to our overall top ten players, and Ken Jennings ranks first at 160 daily double clues.

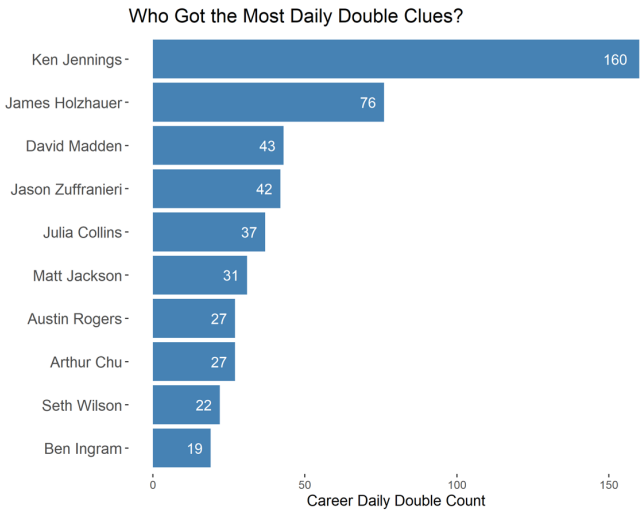

Figure 6: Plot of career Daily Double counts for the top ten players.

V. RShiny App

The analysis of our data is stored in an R Shiny Application. Two interactive charts allow the user to explore the runs of some of Jeopardy!'s most famous and popular players. In addition, the app contains tabs where you can test your knowledge on certain categories of clues, and search through all of the past players. Because our SQL database was not stored in the cloud, the query results were downloaded locally in order to ensure the app's availability to other users.

VI. Challenges & Lessons Learned

Throughout the course of our project, we encountered many challenges in collecting and extracting our data from the J! Archive, manipulating the data into a tidy format, creating our database, and creating our app. We will outline a few of those challenges as well as potential future endeavors below.

A. Data Extraction

Upon discovering the package needed to extract the data, our team uncovered that it output the data in a series of embedded lists. To tidy the data, we created a function to unlist the data and add a unique game identifier. This allowed us to begin structuring the database.

B. Database Creation

Once the database structure started development, we ran into issues normalizing the design because of the player-related data within the location variable. The variable was not reported in a uniform method, so we could not separate out the player's location into city, state, or country. Although that discrepancy occurred with our normalization of the database, we were ready to import all the data.

We used MySQL Workbench's Import Wizard functionality to get the data into the database, although we ran into several challenges. These included matching the definitions of columns to the data, defining foreign keys, uploading the data in the proper sequence so as fit our foreign key definitions, and processing special characters in the data.

Our team effectively overcame each obstacle faced. Through the project development and challenges, we were able to apply what was taught in this course to a data set and run into real problems associated with database creation.

C. Shiny App

The majority of the challenges we ran into with the Shiny app can be put down to lack of familiarity with the tools. Getting the data properly fit into the reactive elements of the Shiny framework proved more challenging than we expected.

D. Future Improvements

Having finished this project, we identified three possible areas of future expansion. First would be the hosting of this database in a publicly accessible way for others to access themselves via SQL. Second would be to finish capturing the remaining data fields, primarily around scoring, held on J!Archive. Finally, our methods here were built as static operations; ideally a continuous feed of information between our database and J! Archive could be constructed so the database remains accurate as new games are added.

VII. References

[1] "J! Archive," [Online]. Available: https://j-archive.com. [Accessed 22 03 2021].

[2] K. Q. Seelye, "The New York Times," 08 11 2020. [Online]. Available: https://www.nytimes.com/2020/11/08/arts/television/alex-trebek-dead.html. [Accessed 22 03 2021].

[3] "The Jeopardy! Fan," [Online]. Available: https://thejeopardyfan.com. [Accessed 22 03 2021].

[4] "Jeopardy! History Wiki," [Online]. Available: https://jeopardyhistory.fandom.com/wiki/Jeopardy!_Statistics. [Accessed 22 03 2021].

[5] O. Roeder, "FiveThirtyEight," 30 04 2019. [Online]. Available: https://fivethirtyeight.com/features/the-battle-for-jeopardy-supremacy/. [Accessed 22 03 2021].

[6] M. Farrow, "GitHub," [Online]. Available: https://github.com/mattfarrow1/7330-term-project.

[7] K. Nicholls, "whatr," [Online]. Available: https://kiernann.com/whatr/. [Accessed 22 03 2021].

[8] King World Productions, Inc., "Press Release," King World, 08 09 2003. [Online]. Available: https://web.archive.org/web/20070928190202/http:/www.kingworld.com/PressRelease.aspx?pressReleaseID=126. [Accessed 27 03 2021].

[9] Jeopardy!, "J!Buzz," Jeopardy!, 07 10 2016. [Online]. Available: https://www.jeopardy.com/jbuzz/behind-scenes/5-jeopardy-rules-every-contestant-should-know. [Accessed 29 03 2021].

[10] R. Kalland, "'Jeopardy!' Champ James Holzhauer's Daily Double Dominance Is The Evolution Of Recent Strategy," 25 04 2019. [Online]. Available: https://uproxx.com/sports/jeopardy-james-holzhauer-april-25-daily-doubles-strategy-historic-wagers/. [Accessed 22 03 2021].

[11] C. Jacobs, "The Federalist," 15 08 2020. [Online]. Available: https://thefederalist.com/2020/08/15/how-alex-trebeks-memoir-explains-the-enduring-success-of-jeopardy/ . [Accessed 22 03 2021].

[12] C. McNear and K. Jennings, Answers in the Form of Questions, New York: Twelve, 2020.