[[1]](#footnote-1)

Analyzing *Jeopardy!* Data

M. Ball, M. Farrow, J. Harrison

*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 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 will be able to connect the database with an RShiny application.

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

# INTRODUCTION

J

eopardy! 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. ABOUT THE SHOW’S POPULARITY [2]

Although the original version of *Jeopardy!* premiered in 1964, the show’s revival in 1984 with host Alex Trebek has led to a run of more than 8,000 games that we wanted to analyze using the data provided on J! Archive to examine what has made this show such a game show staple.

Using data from 6,775 of those games, we examined trends among game categories, and player appearances, and utilized the database to create an interactive application.

# Existing Research

Jeopardy is a game show that lends itself to trends and statistical studies. Sites such as The Jeopardy! Fan [3] and the Jeopardy! History Wiki [4] provide 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 [5].

While we have no reason to doubt the accuracy of these site’s work, it is currently difficult to reproduce their analyses. A large part of what we want to do with this paper is create a database that a user can mine for their own analysis.

# Data

Our approach to this project involved crawling the J! Archive, 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].

## Data Collection

The whatr [5] 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.

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

Diagram

Description automatically generated

Figure 1: EER Diagram

# 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.

## Categories

One of the first questions that we asked was, “Are there common categories that appear in multiple games?”

Chart

Description automatically generated

Figure 2: Most common categories

## Daily Doubles

A Daily Double in *Jeopardy!* is a clue that allows the player who selected the clue has the opportunity to wager an amount of money, from $5 up to their current score, before seeing the answer. 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.

Contestants like James Holzhauser have built their appearance strategy around actively seeking out Daily Double clues [2].

Insert top Daily Double categories here

An analysis of Daily Double locations revealed a strong placement trend, particularly along the fourth row that more strategic players may seek out.

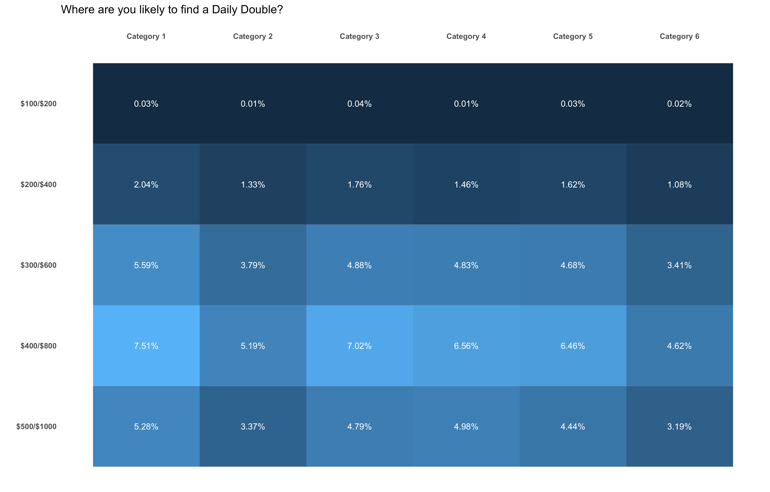


Figure 3: Plot of Daily Double locations

Text

Description automatically generated

Figure : Daily Double wordcloud

With 75 game appearances, Ken Jennings is not only one of the most well-known players, he also got the most Daily Double clues during his time on the show.

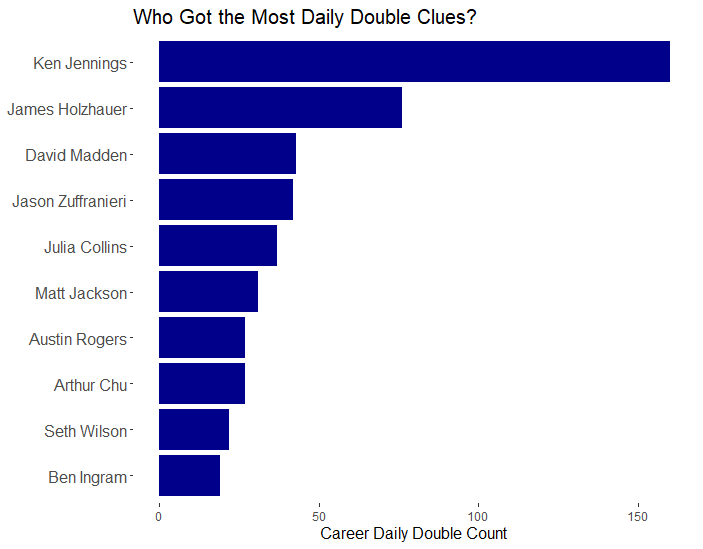


Figure 5: Plot of career Daily Double counts for the top ten players.

## Player Statistics

Following Ken Jennings’ historic run

## Notable 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

# RShiny App

Should I insert a time trend of any notable info for specific players?

Table of worst games?

# References

|  |  |
| --- | --- |
| [1] | "J! Archive," [Online]. Available: https://j-archive.com. [Accessed 22 03 2021]. |
| [2] | 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]. |
| [3] | 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]. |
| [4] | C. McNear and K. Jennings, Answers in the Form of Questions, New York: Twelve, 2020. |

1. 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 \_\_\_\_\_Austin, TX 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 (e-mail: harrisonjp@mail.smu.edu) [↑](#footnote-ref-1)