9-2020

A Data Exploration of Jeopardy! from 1984 to the Present

Brian S. Hamilton
The Graduate Center, City University of New York

A DATA EXPLORATION OF JEOPARDY! FROM 1984 TO THE PRESENT

by

BRIAN HAMILTON

A master's capstone submitted to the Graduate Faculty in Liberal Studies in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2020

A Data Exploration of Jeopardy! From 1984 to the Present

by

Brian Hamilton

This manuscript has been read and accepted for the Graduate Faculty in Liberal Studies in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Matthew K. Gold Thesis Advisor
Date	Elizabeth Macaulay-Lewis Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

A Data Exploration of Jeopardy! From 1984 to the Present

by

Brian Hamilton

Advisor: Matthew K. Gold

The gameshow *Jeopardy!* has been around in its current iteration—hosted by Alex Trebek—since 1984. During this time, it has accumulated data on clues, contestants, and possible strategies on how to win. Using a crowd-sourced archive called *J! Archive*, this project seeks to find trends in the topics that the game covers and take a deeper look into the performance of its contestants. It employs topic modeling, a text-analysis method, to organize the hundreds of thousands of archived clues and statistical analysis to rate the performance of contestants by gender. Using web-based visualization tools, the data is shown in an interactive and understandable way. The main goal of this project is to take a dive into the data, analyze a series of points, create a robust database for the show, and connect *Jeopardy!* to a larger cultural, social, and political context. This will allow for further analysis and visualizations to be done in the future.

TABLE OF CONTENTS

List of Figures……....vi

Digital Manifest……vii

Note on Technical Specifications……viii

1. Introduction……1

2. Historical Analysis and Research……5

3. Development and Visualizations……...10

4. Continuation……..17

Works Cited……...19

LIST OF FIGURES

Figure 1: Entity-Relationship Diagram……...4

Figure 2: Shortened Topic Models……12

Figure 3: Chi-Square Test of Independence for Gender and Religion Clues……....16

DIGITAL MANIFEST

1. **Capstone Whitepaper (PDF)**

2. **WARC Files**

    a. **Project Website**
    Archived version of https://jeopardy.brianhamilton.me

3. **Code and Other Deliverables**

    a. **Database (SQL Source File)**
    Backup of the PostgreSQL database

    b. **Zip File for GitHub Repositories**
        i. https://github.com/hmltnbrn/jeopardy-scrape
        Python scripts for scraping *J! Archive*

        ii. https://github.com/hmltnbrn/jeopardy-viz
        HTML/CSS/JavaScript files for project website

NOTE ON TECHNICAL SPECIFICATIONS

1. Python Scraping Scripts

In order to run the scripts, Python 3.7 and the Psycopg2 and BeautifulSoup packages are required. All commands should be run in a command prompt. A PostgreSQL database needs to be set up and a credentials.json file should be placed in the credentials directory. It should include information about the host, database name, username, and password. A SQL creation script is present in the database directory. This will create all necessary database tables. To see descriptions for each possible command, run the command below:

```
cd scrape
python scrape.py –help
```

More information on how to run the topic modeling scripts can be found in the readme of the GitHub repository (https://github.com/hmltnbrn/jeopardy-scrape) or the methods pages of the website (https://jeopardy.brianhamilton.me/methods).

2. JavaScript Website

Running the website locally requires a systemwide installation of Node.js. All packages can then be installed by running the command below:

```
yarn install
```

Run the following command to compile the site:

```
yarn run dev
```

Navigate to http://localhost:3000 to see the site.

1. Introduction

In 1984, the fourth iteration of the daily syndicated game show *Jeopardy!* was first hosted by Alex Trebek. Nearly forty years later, it still runs daily, and is considered to be one of the most popular game shows on the air. The show has accumulated dozens of awards—including some for its host—and is routinely ranked as one of the greatest game shows of all time. Some of the show's more prolific champions, such as Ken Jennings and James Holzhauer, have become household names. Broadcast game shows and trivia shows in general have been part of American culture since at least the 1930s, when radio show variations began to air (Hoerschelmann).

The *Jeopardy!* game show was initially created in 1964 by Merv Griffin and hosted by Art Fleming in three different iterations until 1979. The format has always been the same—two rounds with six categories, each with five different clues, and a final round with a single clue of a random category. The second round has double the number of dollars given for a correct answer as the first round. Since 1964, the dollar amounts have periodically been raised to account for inflation. This was most recently done in 2001. The final round is different, allowing each contestant to wager an amount of money after seeing the category and before seeing the clue. At the end of the game, the contestant with the highest amount of money wins and continues on to face a new set of two opponents in the next episode.

The idea of a project[1] based around analyzing *Jeopardy!* grew out of an interest in trivia that I have had for most of my teenage and adult life. I used to watch my parents play along with the show every night and then when I grew older, I found an interest in bar trivia nights and watching *Jeopardy!* myself. Even though there are definitely some subjects I am not good at and I

[1] https://jeopardy.brianhamilton.me/

may not be a wizard at trivia as a whole, I have always loved playing.

After I graduated from college in 2014, I moved to New York City, and, due to the nature of people moving apart over time, separated myself from most of the friends I had in college. It was with them that I first got into bar trivia. Without them, I did not go as much. However, in 2017, I met up with them again on the chat platform Discord. It was there, through one of my friends, that I was introduced to *J! Archive*,[2] a crowd-sourced archive of a large amount of *Jeopardy!* episodes, starting in 1984 with Alex Trebek's term as the host. We would go to random shows and quiz each other, keeping a tally of how well each person was doing. After a while playing like this, I started to wonder whether there was a better way to play, rather than manually loading the site and going down a list of clues until one moved onto the next game. As a software engineer, my mind began to run wild with ideas of a programming solution to the problem. From that, I created Trebot.[3]

Trebot is a JavaScript Discord bot that takes commands and gives players random *Jeopardy!* questions and eventually their answers. In order to get what I needed to populate the database, I had to write a Python script that scraped *J! Archive* of the relevant data. Unfortunately, there was not an easy way to accumulate the data from *J! Archive* and I was forced to resort to scraping—reading the underlying code of a website and parsing it to find something relevant. From all of the information present on the site, I only needed the data relevant to the trivia bot I was creating. Thus, all I took were clues, answers, and the dollar value for each. Information on contestants, right or wrong guesses, or any other extraneous data was not important. Eventually, I was able to get a clean database of everything I needed on *J! Archive* and stored it on the cloud using MongoDB. Now, instead of manually compiling questions, my friends and I could call a

[2] http://j-archive.com/
[3] https://github.com/hmltnbrn/trebot

command from the bot and it would supply us with a random question from any time in *Jeopardy!*'s history. As the years went on, I would add more features; the ability to keep score and string similarity algorithms to determine whether a user's answer was correct or not.

For most of my time in the MALS program at the Graduate Center, I was considering projects that I could work on for my capstone. Concentrating in data visualization, I always figured that it would have to be related in some way to data. In the spring of 2018, I took the MALS course on data visualization. During one of the classes, after being shown a few examples of visualization projects, I came up with the idea to expand upon the scraping script that I had written before and create an even larger database of information that was present on *J! Archive*. Now, instead of just a simple database of clues and answers, I could expand it to be relational, allowing different data points to connect to one another. I still used Python to scrape the site, but the script ended up being much more intricate.

J! Archive is not an official database, but is instead a crowd-sourced archive by fans and former contestants. Crowdsourcing has the advantage of having "the ability to reach and engage a broader intelligence pool" (Milo 64). In this case, the broader pool is dedicated fans of *Jeopardy!*. Created and initially maintained by previous *Jeopardy!* contestant Robert Knecht Schmidt, it now, according to Schmidt, "kind of lives on its own," with multiple archivists transcribing each episode and uploading the data to the archive daily (D'Addario 2011). As of April of 2020, with the show in the midst of its thirty-sixth season, the archive boasts of containing over 380,000 clues—adding more each day an episode airs on television. Alex Trebek, when told of the archive in an interview in 2011, remarked that the fans who maintain it should "get a life" (Stone 2011). However, some future contestants use the archive as a way to practice for their appearances, making the site a useful tool for them and a source of enjoyment for some who enjoy engaging in trivia.

In order to access the data present on the site, I created a Python script designed to scrape each relevant page. This is because the site does not contain an API that would allow me to access the data easily. I made the script available as an open source resource.[4] A PostgreSQL relational database is used to store the data. This allows each contestant, category, and episode to interact with each other. For example, one contestant was allowed to be part of multiple episodes and a category could be matched to multiple clues. *Figure 1* shows an entity-relationship diagram of the database. I decided against using a cloud-based database, like Amazon Web Services, due to the monetary cost associated with it. Instead, I opted to store the data locally on my personal desktop. This can be dangerous, due to a potential system failure, but with periodic backups done, that should not be a problem.

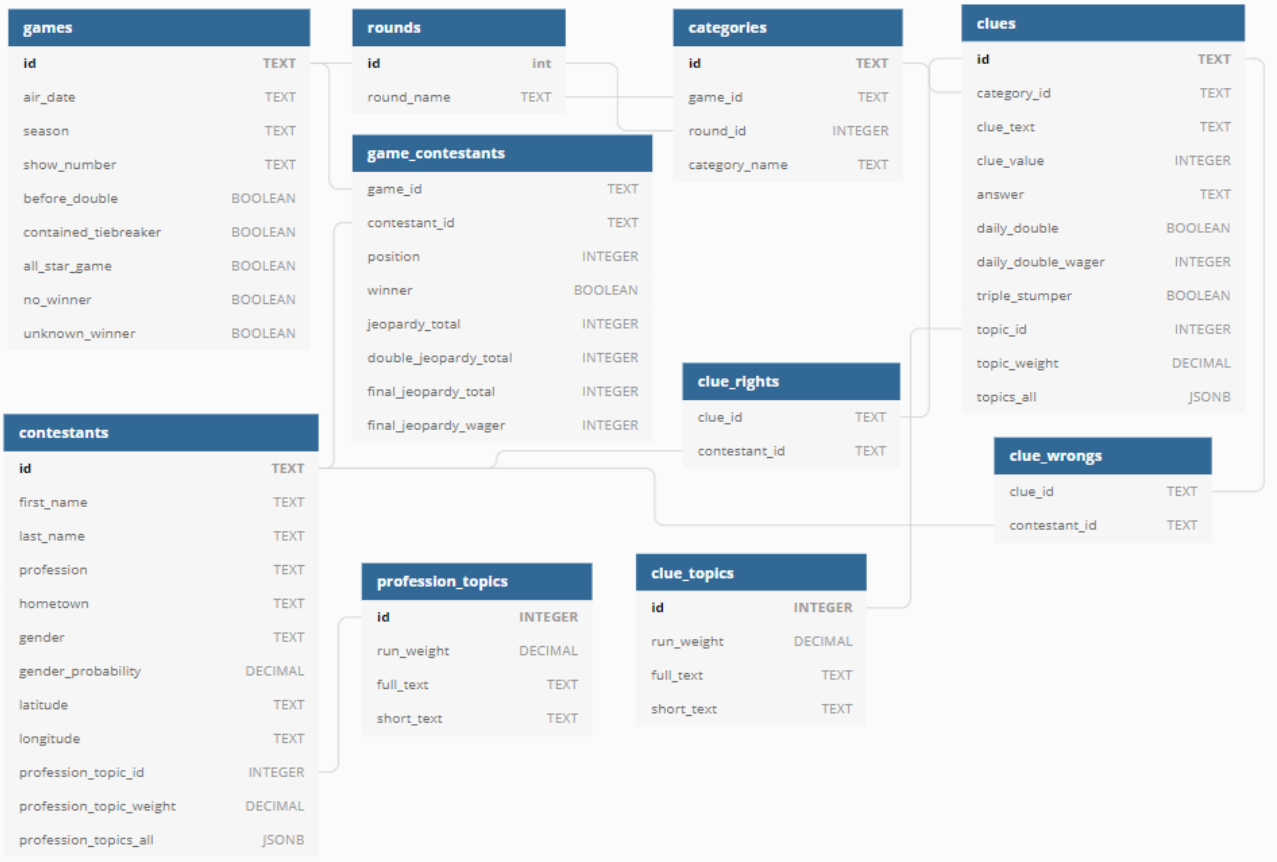

Figure 1: Entity-Relationship Diagram

[4] https://github.com/hmltnbrn/jeopardy-scrape

The script scans for the appropriate pages (seasons and then episodes) and then makes HTTP[5] calls to the site in order to retrieve the HTML[6] for each individual episode available on the site. After this was done, the raw HTML is parsed, finding the exact data I was looking for. In some cases, there would be instances where an episode's page would have a small difference. These differences could include missing data or a website design mistake. After running the script through, I would have to account for this by using if/else statements. Episode parity was another major change that needed to be considered. Some episodes had tiebreakers and some occurred before the money amounts were doubled. Some of these differences were given their own field in the database as Boolean[7] values. Once this parsing was complete, the completed data object for each episode was inserted into the database. From this database, I would be able to begin my analysis.

Jeopardy! has been present throughout the highs and lows of American history and culture over the last forty years. This history can be found within its clues, categories, answers, and contestants. Current events, movies, celebrities, wars, politicians, cultural trends—it is all there within the game itself. Taking a snapshot of a period of *Jeopardy!* history can reveal details of the time it was made. At the same time, this relationship can be seen from the other direction: how do the producers and writers view the world around them? The clues and categories that they contribute to the show can tell us what they believe is most important to the world and what kind of bias could be present in the show's writers' questions. In the end, the basic aim of this project is to see what can be found in the data.

2. Historical Analysis and Research

[5] Hypertext Transfer Protocol—communication protocol used for accessing website data.
[6] Hypertext Markup Language—programming language used to build the front-end for a website.
[7] A true or false variable.

Jeopardy! is a game rooted in basic trivial knowledge, but also mathematics, economics, and behavior. Most literature in relation to *Jeopardy!* ties into that. Prior to *J! Archive*, research could only use a subset of data from the show's history. Now, with the help of outside contributors through crowdsourcing, a researcher could expand that to include most of the show's history.

Broadcast game shows have been around for almost 100 years, starting on the radio and eventually moving to television (Hoerschelmann). *Jeopardy!* itself has been around since 1964. However, it is important to understand how the show works behind the scenes and where these questions come from. Generally, the writers try to use two different sources for each question—using publications like *Encyclopedia Americana*, *Encyclopedia Brittanica*, and the *Oxford Companion to English Literature* before the widespread use of electronic sources (Berthold 12). In the modern era, *Jeopardy!* relies on, along with its old sources, Oxford University online databases (Roncevic). Most questions are generally considered to be basic, with only a surface-level understanding of the topic needed to be able to answer them. It would be rare to find obscure topics in a question. According to Kathy Easterling, a former *Jeopardy!* writer, "Our audience must relate to the material in some way. Our producer wants the audience to relate to every question in some way—even if they haven't heard of Mark Twain's book about Joan of Arc they've heard of both Twain and Joan so we can use a question about it" (Berthold 13).

However, does the use of these encyclopedias and Oxford online databases in the construction of questions, along with the basic level knowledge needed for answering them, reveal a bias in the way they are written? Most topics are going to be centered around the most popular and widely-known parts. For example, authors of the most popular American literature tend to be white men and it can be rare to find references to women or other minority authors (Berthold 14-15). This leads me to believe that *Jeopardy!* is in fact tied to American culture in a very specific

way—namely one that just parrots what most people are taught in grade school and what is going on in the country at the specific time of the episode's airing. These are the things that the audience is more likely to relate to, as Kathy Easterling argues. It is also possible that the biases of the *Jeopardy!* staff itself determines what they believe to be common knowledge. A less diverse writers' room can lead to a segment of trivia knowledge never considered for the show, as the writers are not aware of it.

During this project, I employed the use of natural language processing and machine learning to organize *Jeopardy!* clues into a set of topics. This is not the first time someone has used *Jeopardy!* as a way to parse language. In 2011, the IBM computer Watson competed against Ken Jennings and Brad Rutter, two previous *Jeopardy!* champions (Kroeker 13). In order to get to the answer, Watson used natural language processing, a series of algorithms, and a large database to parse the clue and reach an answer (Rachlin 1). It also used analytics to determine wagers. It was able to do this in a very short time. While I do not have the computing power or computer science knowledge to do something similar, it is interesting to note that past developers have believed that *Jeopardy!* offered a useful source for training language algorithms and employing machine learning.

Watson showed that wagering in *Jeopardy!*'s Final Jeopardy round can also be done through the use of an algorithm. For the most part, contestants use more precise numbers, but Watson opted to use imprecise numbers, as the algorithm calculated (Kroeker 15). It is possible to create these algorithms, as noted by Jeffrey Floyd in "A Discrete Analysis of 'Final Jeopardy.'" A mathematician can consider the rules of the game and create an optimization algorithm that takes into account the many possibilities that can occur leading up to the final round. If employed by contestants, mathematicians believe that these strategies can actually help a contestant make more

money and be more successful in the long run (Gilbert 11).

While some may be able to follow these strategies, it can be difficult to put it into practice. Patrick Headley is a mathematician who competed on the show. He attempted to use a game theory-like strategy, but failed to anticipate what the other contestants would do (Headley 28). Because of this, he lost. This aligns with the findings of Andrew Metrick, an economist who analyzed the behavior of *Jeopardy!* contestants. He found that "many players overestimate their abilities or fail to notice that a specific option is available" (Metrick 252). These options can range from picking a different category that applies to their own strengths or wagering safely. An analysis of wagering and behavior is something I might be able to look at using my *J! Archive* database. While I will not focus on that in this current project, it is definitely something that can be looked at in the future.

A few studies were done based around the differences in gender between *Jeopardy!* contestants. The first is a 1998 study done by Sheila Brownlow, Rebecca Whitener, and Janet Rupert of Catawba College titled "I'll Take Gender Differences for $1,000!" The idea behind the study was to find out whether men and women answered topics differently, responding correctly or incorrectly to gendered topics. The paper states that "women consistently underestimate their abilities and are often inaccurate in making judgements about how capable they are, particularly when the task at hand is considered to be typically masculine" (271).

The writers of the paper watched sixty-five *Jeopardy!* episodes televised in the first half of 1996 and kept track of the questions and contestants (272). Using an outside testing group of fifteen men and women, they took the questions asked in each of those episodes and manually generated a map of "femininity," as in if the question was more likely to be known by men or women (273). The contestants of this half-year sample of *Jeopardy!* were made up of 59.5% men

and 40.5% women (274). Comparing this data to the performance of the contestants on the actual *Jeopardy!* episodes, they found that women did better in feminine topics than masculine ones (281). Another interesting data point that the study found was that a higher proportion of women won one game and a higher proportion of men won more than one game (277). This study offers a good departure point for analysis into gender roles in *Jeopardy!* topics, some of which I can attempt to back up with my own data.

The second gender-based study is the 2013 paper "Gender in Jeopardy!" by Thomas Linneman. It focuses primarily on the issue of uptalk, the act of raising your voice in a questioning manner near the end of a statement. The author notes that in sociolinguistic literature, it is generally believed that uptalk is more common in women, especially young white women (83). This study wanted to take this idea and put to it to the test under different circumstances during a typical *Jeopardy!* episode, including when a contestant was ahead or when there were more men or more women playing. To find the data, Linneman watched 100 episodes from the 2009-2010 season and determined if a contestant employed uptalk (89). In the process of building his demographic data for contestant race and age, Linneman actually employed the use of *J! Archive* and the images of the contestants present on the site (89).

While a study on uptalk is a good use of a continually evolving game show, it is not something that I can work with in my own *J! Archive* data. Unfortunately, due to not actually being able to see the episode itself, I would not be able to discern a speaker's pitch or tone. While this is disappointing, there are a few data points that Linneman found that would be useful for me to look at. Of the 300 contestants that the study focused on, 56% were men and 44% were women (91). Despite being a small sample size, this should be representative of the actual gender breakdown of the show and my dataset can prove that. The study also noted the breakdown of contestants in an

episode, namely how many were men and how many were women. According to Linneman, "the most common contestant configuration on the show was two men and one woman, followed by two women and one man, and finally three men (there were no episodes with three women)" (91). I can group the data in a way to see if this configuration holds up over the course of the show's history, and how many episodes actually have three women.

3. Development and Visualization

After I finished collecting the data from *J! Archive* and completed my research into past projects, I began to look for ideas on how to analyze it and build the data visualizations. The first thing I had to do was set up the infrastructure required for the website. Along with the use of basic HTML and CSS, I opted to use a JavaScript; Node.js[8] for the back-end and D3.js[9] for the front-end. Node.js and D3.js are JavaScript-based frameworks for building a web server and interactive data visualizations, respectively. I have considerable experience with both in my career and I believed that they would be the best tools to use for a project like this. I hosted the project[10] on my own personal domain and made the code for both the website[11] and the scraping[12] available on GitHub.

The first type of analysis I did was to use topic modeling. Topic modeling is the act of running a corpus of text through an algorithm, finding the words that have a high probability of occurring together (Steyvers). I wanted to create a topic model of the roughly 375,000 clues that I was able to scrape from the site—separating each out into twenty-five topics. I would do this by using MALLET,[13] a Java-based command line software package. This package uses a series of

[8] https://nodejs.org/en/
[9] https://d3js.org/
[10] https://jeopardy.brianhamilton.me/
[11] https://github.com/hmltnbrn/jeopardy-viz
[12] https://github.com/hmltnbrn/jeopardy-scrape
[13] http://mallet.cs.umass.edu/

algorithms and machine learning to build this topic model. My main goal of this was to narrow down what *Jeopardy!* questions focus on, giving me a better idea of what most are generally about. I wrote a series of Python scripts that built the necessary data files that would be used by MALLET in the process of building the model, as well as scripts that would insert the topic model into the database. These are also present in the GitHub repository. Once the model was completed and the series of topics was in the database, I shortened the topic keywords to what I believed made the most sense, as shown in *Figure 2*. The full text is a long string of words that the model put together, with the ones earlier in the string more likely to appear in that specific topic. The short text was my best guess and could easily be taken in a different direction by someone else.

Figure 2: Shortened Topic Models

Full Text	Short Text
ancient god bible greek religion biblical b.c roman book king mythology man jesus testament people day history called son religious	Religion
science it's element type water gas light weather metal physics called energy elements earth air chemistry measures rock made measure	Science
art artists fashion artist painting french century style painted work made named famous hat museum wear worn paintings called painter	Art
city museum york building san street world london capital house city's home park built hotel bridge it's tower museums famous	Geography/Local
country world island capital south sea city islands river geography it's countries largest africa north nation country's water african miles	Geography/World
state city u.s river states national lake capital named park cities south north it's west california american state's texas county	Geography/U.S.
book wrote author authors literature title books literary american story john lit novels man poetry poet published century character poem	Literature
show series played sitcom family married television star actress title woman women famous shows celebrity character host i'm film comic	Television
school university college prize colleges nobel magazine founded york times universities newspaper named won news magazines awards high newspapers american	Academics
food it's made cheese dish french cooking type eat cream drink meat make soup bread candy sauce chocolate chicken called	Food & Drink
u.s law government money court it's act organization department american history group united rights crime legal states type bill amendment	Law
company business car brand industry stock introduced million store made u.s cars company's sold ford model founded it's product names	Business
clue crew sarah jimmy reports space shows kelly monitor it's presents moon delivers earth planet cheryl star math science sun	Random/Clue Crew
body medicine blood it's disease human medical part health heart called type brain organ science anatomy eye bones bone doctor	Human Body
animal animals bird dog birds fish it's type species called dogs cat sea named bear mammals breed feet creature wild	Animals
song music hit songs rock musical band pop love singer album top title group country dance heard it's sang instrument	Music/General
film movie movies played title oscar films star actor man role won actors john character classic series big roles screen	Film
opera play shakespeare music ballet title classical wrote composer composers musical characters plays theatre character shakespearean tale based literary dance	Music/Classical
king history country world queen century british leader years minister great prime prince empire born man france president england french	History/General
president u.s presidential john day state house presidents man george american secretary senator vice history lincoln governor party washington chief	U.S. Politics
war u.s battle american history civil world military army general british ship captain john century air navy revolution historic ships	History/War
word words letter it's means meaning term latin clues crossword phrase phrases french you're greek adjective letters person time origins	Words
sports team baseball football won sport olympic world game league games hall record bowl home olympics nfl fame player college	Sports
it's game type words time games called letter term ball you're play word board feet back piece hand make horse	Random/Games
color tree red it's fruit plant drink wine white flower green type flowers blue trees beer colors made potent plants	Random/Colors

With the model shortened, I grouped the topics into a time series that showed the amount by percentage that each was used in the time span of 1984-2019. My hope from this visualization was to see trends. I believe that even minor changes in the proportion of usage for a topic can be significant. From the start, I wanted to know if a topic about war was more prevalent during times of heightened American foreign conflict; as in, during the Gulf War or the Iraq War. From the visualization shown on the site, there is a noticeable increase in war-based questions in the early 1990s, potentially backing up that idea. At the same time, the topic model algorithm saw a link with "Iraq" and "war," due to the number of times it was seen together. This is called colocation. Even a question about the something like the geography or basic history of Iraq would be classified as a "war" question.

There is one major drawback of using a topic modeling algorithm in an attempt to classify each question. The model will have false positives like the Iraq War colocation I specified before. One other egregious example I found was an obvious music question about the band Green Day being associated with the sports topic. However, I believe they will not happen enough times to be significant, as most Green Day questions are classified in the music topic, as they should be. The best way I found to reduce these errors is by including the name of the category in the model, along with the question and answer. By category, I mean the one that *Jeopardy!* itself uses to put five clues together during a game. While the show does not always use descriptive categories and often relies on puns or other unrelated words, it would be helpful to the algorithm to take it into consideration. I made this choice after noticing the false positives from only using the question and answer.

Contestant information from *J! Archive* includes their first and last name, hometown, and profession. For American contestants, the hometown specifies the city and U.S. state they come

from. I grouped the data in a way to count the number of contestants representing each of the fifty states and the District of Columbia. I hoped to see a noticeable difference between the small and large states. Alone, the number of contestants would not be useful, as it would only tell me that larger states have more contestants—in my opinion, an obvious conclusion. Instead, I used 2019 population estimates from the United States Census Bureau to normalize the results, giving me a proportion of the population that has been a contestant on *Jeopardy!*. In can be seen, that even with normalizing for population, larger states are more likely to be represented, with Texas and the District of Columbia being the noticeable outsiders; Texas is underrepresented and the District of Columbia is overrepresented. A possible explanation for the overrepresentation of the District of Columbia is that some episodes of *Jeopardy!* are filmed there, usually having local political journalists and pundits.

A major component of past research studies into *Jeopardy!* focused on the gender imbalance on the show. I wanted to do the same, in order to both re-analyze what these studies found and make new discoveries. While contestant information from *J! Archive* includes their first and last name, it does not include a direct reference to their preferred gender. In order to classify contestants in this way, I used a public API called Genderize.io[14] to generate the likely gender of each contestant that had a valid first name. The Python script I wrote would call the API, get a return value of their likely gender, and update the database. With this new classification of each contestant, I could do even more analysis of the data.[15]

The first two visualizations that I created were meant to display the number of men and

[14] https://genderize.io/

[15] Using binary gender as a classification can be seen as problematic, especially when using an algorithm to determine it, with no input from the contestant. As argued by Laura Mandell, "rewriting gender as sex, wrongly understood as a simple binary opposition, transforms it into a weak signal that might be statistically "correct"" (17). In this project, binary gender will be used for analysis, as the overall breakdown of contestants by gender is a good starting point. However, for future iterations, separating contestants into more categories could be the better option.

women that have played during the game's history. I was able to group the data in two ways; the first by the total number of men and women that have played and the second by the breakdown of the three contestants in a game. This analysis was able to back up data found in "I'll Take Gender Differences for $1,000!" and "Gender in Jeopardy!," Those studies used a subset of episodes, rather than the entire history of the show. I found that 42.5% of contestants were women and 57.5% were men, coming very close to the percentages found in those studies. I was also able to back up the breakdown of the three contestants, using games in which all three had a specified gender. According to "Gender in Jeopardy!," two men and one woman was the most likely setup of contestants. The study also found that in its own subset of games, there was never an instance of three women competing against each other. My own analysis of the data saw that this actually happened 116 times over the show's history.

In my initial collection of the data from *J! Archive*, I found which clues individual contestants got right or wrong and stored the data in a separate database table. Using the topic modeling I created and the gender breakdown, I did a Chi-Square Test of Independence for each topic. This test determines if variables—in this case, gender and correctness of the clue—are related or completely independent. I had to do a more complex analysis like this because the male contestants outnumber female contestants and looking at raw numbers would not tell the full story. *Figure 3* shows an example of the test, using the topic of religion. A chi state value with a number higher than the critical value is significant and the categories are considered related. My goal in doing this analysis was to find if any topics were gendered—if a topic is more likely to be known by men or women. For example, sports clues were significant for men and food/drink clues were significant toward women. This finding is consistent with that of "I'll Take Gender Differences for $1,000!", showing that men and women did perform differently with certain topics.

Religion			
Observed	Correct	Incorrect	Grand Total
Male	8814	1813	10627
Female	4713	1033	5746
Total	13527	2846	16373
Expected	Correct	Incorrect	
Male	8779.785562	1847.214438	
Female	4747.214438	998.7855616	
Chi-square	Correct	Incorrect	
Male	0.133332163	0.633725988	
Female	0.246592567	1.172051179	
Chi state	2.185701897		
df	1		
Critical	3.841458821		
p-value	0.139297548		

Figure 3: Chi-Square Test of Independence for Gender and Religion Clues

To continue the analysis of gender and *Jeopardy!*, I did a breakdown of wins between men and women. I grouped each contestant into three sub-groups: zero wins, one win, and more than one win. A stacked bar chart visualization shows the relationship. What I found is that, in comparison to men, a smaller percentage of women were likely to win one game and more than one game. This finding actually contradicts "I'll Take Gender Differences for $1,000!." In that study's sample of episodes, a higher percentage of women won only one game. However, my own data, covering most of the show's history, said the opposite.

The last visualization I created was another basic time series showing the final total for the winning contestant. Grouping by year and gender, I took the median amount to determine a how much was roughly won each game. The three most obvious points of reference are 2001-2002, 2004, and 2019. Between 2001 and 2002, the dollar amounts for each clue were doubled. During 2004 and 2019, two of the most successful contestants competed; Ken Jennings in 2004 and James

Holzhauer in 2019. James Holzhauer was known for winning by a large margin, racking up a lot of money each game. 2011 and 2019 also saw an increase in median winnings for women. However, while I looked for a specific contestant that could've caused this, there were no long-term female winners in those years. The most successful female contestant, Julia Collins, competed in 2014--a year with no major change.

4. Continuation

One of the original big ideas for this project was an interactive visualization that allowed for the user to query the database themselves. What would happen is that the user would input a line of text and a search for it would be done inside the database in the categories, clues, and answers. Results for it would then be shown on a time series. This would allow for a user to see when a topic was covered, possible showing trends. To do this, I would have needed to host the database on the cloud and build out a back-end infrastructure that included API calls. That was not something I wanted to do for this, but would be one of the first things I build when continuing this project.

The database that I created can be a good start for running more topic modeling analysis. Prior to starting the project, I considered building a topic model on the professions that I scraped from *J! Archive*. I went as far as adding the database table and appropriate columns for it. However, I ran into the issue of the modeling algorithm not creating distinct topics. It had too much variation between the runs, showing either duplicates or missing potential topics entirely. There would need to be more testing to determine what the best course of action is for modeling professions. I was not able to come to a decision on that for this project and it can be something that I work on later on.

There is still a great deal of analysis that can be done with *Jeopardy!* and I am excited to continue working with it. I think this project will also be useful to other academics who are looking to do more research in understanding the cultural impact of *Jeopardy!* and how contestants have fared over time, using my findings and those of the literature I found as a base. I plan on making a backup of my database available publicly for those who are interested.

Works Cited

Berthold, Michael C. “Jeopardy!, Cultural Literacy, and the Discourse of Trivia.” *The Journal of American Culture*, vol. 13, no. 1, 1 Mar. 1990, pp. 11–17., doi:10.1111/j.1542-734x.1990.1301_11.x.

Brown, Scott. *Social Information: Gaining Competitive and Business Advantage Using Social Media Tools*, Elsevier Science & Technology, 2012. ProQuest Ebook Central.

Brownlow, S., Whitener, R. & Rupert, J.M. “I'll Take Gender Differences for $1000!” Domain-Specific Intellectual Success on “Jeopardy”. *Sex Roles* 38, 269–285 (1998). https://doi-org.ezproxy.gc.cuny.edu/10.1023/A:1018789201377.

D'Addario, Daniel. “Inside J-Archive, the Nearly Comprehensive Online Jeopardy! Archive Maintained by Obsessive Fans.” Slate Magazine, Slate, 11 Feb. 2011.

Floyd, Jeffrey K. “A Discrete Analysis of ‘Final Jeopardy.’” *The Mathematics Teacher*, vol. 87, no. 5, 1994, pp. 328–331.

Gilbert, George T., and Rhonda L. Hatcher. “Wagering in Final Jeopardy*!” Mathematics Magazine*, vol. 67, no. 4, 1994, pp. 268–277.

Headley, Patrick. “How I Lost on Jeopardy!” *Math Horizons*, vol. 6, no. 4, Apr. 1999, pp. 27–28., doi:10.1080/10724117.1999.11975104.

Hoerschelmann, Olaf. *Rules of the Game: Quiz Shows and American Culture*. State University of New York Press, 2006.

Kroeker, Kirk L. “Weighing Watson's Impact.(IBM Watson Computer).” *Communications of the ACM*, vol. 54, no. 7, 2011, pp. 13–15.

Linneman, Thomas J. “Gender in Jeopardy!” *Gender & Society*, vol. 27, no. 1, 2012, pp. 82–105., doi:10.1177/0891243212464905.

Mandell, Laura. “Gender and Cultural Analytics: Finding or Making Stereotypes?” *Debates in the*

Digital Humanities 2019, edited by Matthew K. Gold and Lauren F. Klein, University of Minnesota Press, Minneapolis; London, 2019, pp. 3-26.

Metrick, Andrew. “A Natural Experiment in ‘Jeopardy!".” *The American Economic Review*, vol. 85, no. 1, 1995, pp. 240–253.

Milo, Tova. “Crowd-Based Data Sourcing.” *Databases in Networked Information Systems Lecture Notes in Computer Science*, 2011, pp. 64–67., doi:10.1007/978-3-642-25731-5_6.

Rachlin, H. “Making IBM’s Computer, Watson, Human.” *The Behavior Analyst*. 35, 1–16 (2012). https://doi.org/10.1007/BF03392260.

Roncevic, Mirela. "Two Oxford University Press endeavors worth noting: the publisher has just signed an agreement with the producers of the game show Jeopardy! to give their researchers access to a number of Oxford's electronic resources (among them, OED Online, ODNB Online, and ANB Online) to verify the 14,000 questions and answers written for the show each season." *Library Journal*, 1 Feb. 2006, p. 106. Gale Literature Resource Center, https://link-gale-com.ezproxy.gc.cuny.edu/apps/doc/A142297062/LitRC?u=cuny_gradctr&sid=LitRC&xid=dadefac9.

Steyvers, M. and Griffiths, T. “Probabilistic Topic Models.” *Latent Semantic Analysis: A Road to Meaning*, edited by T. Landauer, D. McNamara, S. Dennis, and W. Kintsch. Laurence Erlbaum, 2006.

Stone, Daniel. “Alex Trebek Interview on Jeopardy, Watson, IPads, and Retirement.” The Daily Beast, The Daily Beast Company, 13 Feb. 2011.