**Jeopardy Data Analysis**

Jakub Gierus – Final Project

# Introduction

"Jeopardy!" is a classic American television game show, renowned for its unique answer-and-question format. Conceived by Merv Griffin, the show originally premiered in 1964 and underwent several iterations before settling into its most famous version in 1984, hosted by Alex Trebek until his death in 2020. Contestants on "Jeopardy!" compete to answer questions from various categories, presented in the form of answers to which they must supply the questions. The show is divided into three rounds: the Jeopardy round, the Double Jeopardy round, and Final Jeopardy, where contestants can wager their earnings. The first two rounds include 6 categories, and 5 questions per category, totaling 30 question per round, and a maximum of 61 questions in a game. The winner of a "Jeopardy!" game gets to be on the next game.

The dataset used in this project includes question, episode and contestant data on the 217 episodes of the show's 35th season, from Show #6096 (2018-09-10) to Show #8045 (2019-07-26).

Question: While fans of "Jeopardy!" are often fans of trivia, a lot of them are also big fans of the contestants, especially those that win a lot. Necessarily, by the format of the show, if a contestant wins a lot, they must also be in a streak of wins, since the moment they lose, they are off the show (with rare exceptions). The longer the streak, the more acclaim a contestant gets. Some of the most famous contestants are ones that had very long winning streak, like Ken Jennings (74 game win streak), Amy Schneider (40 game win streak) and James Holzhauer (32 game win streak). In this report, I intend to explore what factors influence the duration of a contestant's win streak, as well as they amount of money they win in a single game.

# Methods

The dataset was acquired from the J-Archive (<https://j-archive.com/>), a database of all Jeopardy questions, shows and contestants. This analysis uses the whatr R package specifically to access the data. The data is cleaned by removing N/A and null values. The wrangling, cleaning descriptions for each specific table will be detailed below.

**Synopses Table**

The synopses table has a game id, first name, final score, right answer number and wrong answer column. For the synopses, I mutate the table to add the correct answer rate (right / (right + wrong)), and a unique id as columns. Additionally, I sort the table by the final score.

**Scores Table**

The scores table has a game id, round (Jeopardy/1, Double Jeopardy/2, Final Jeopardy/3), clue number (i), first name, score and Daily Double boolean. The original score tables only includes contestants that buzz in and attempt the clue. However, during the data wrangling I added an entry for every contestant for every clue, with score being 0 if they didn't buzz in. Then, I calculated the cumulative score for each contestant over the course of each game, and then ranked each game's contestant 1-3.

### Players Table

The players table includes the first and last name of each player. game id, and the contestant description. I wrangled this data, to add the cumulative win streak for each contestant, calculated from the modified scores table.

### Clustering clues

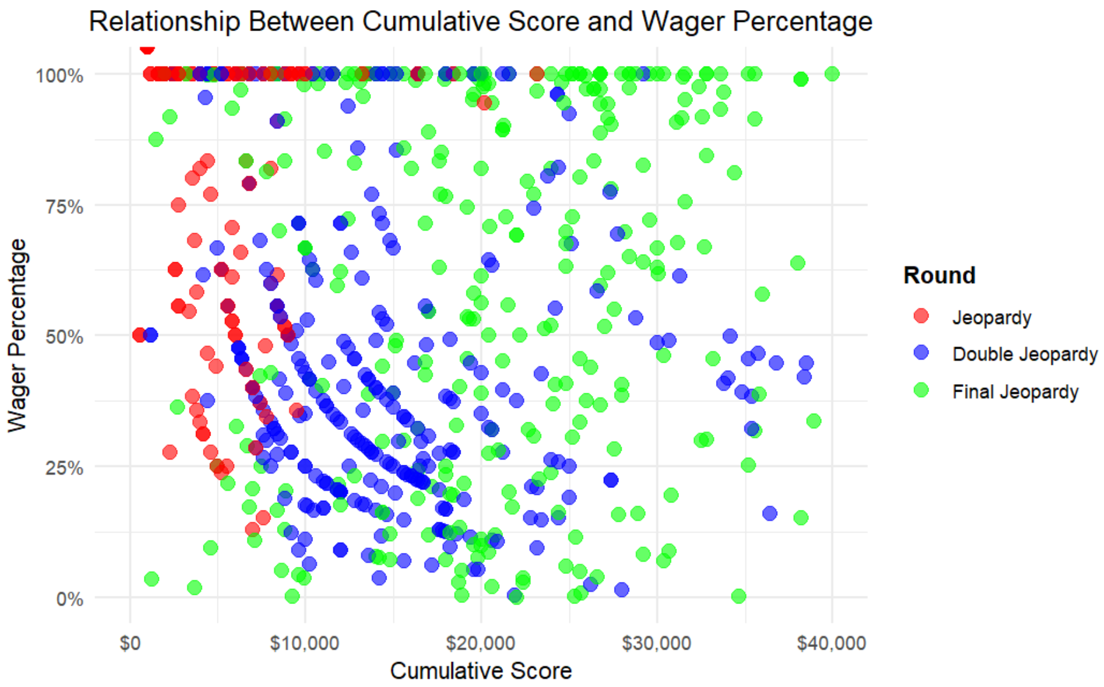
Per Game, there are 12 categories used to categorize five clues each. However, I wanted to analyze effectiveness in relation to areas of knowledge, and wanted to breakdown the unique category names (like “3 Consonants, No Vowels” or “Largest Cities in Asia”) into a small number of archetypal categories.

To identify archetypal categories within the Jeopardy dataset, I employed natural language processing (NLP) techniques and topic modeling. The dataset, consisting of columns for category, clue, answer, and additional metadata, was preprocessed by combining the category, clue, and answer columns into a single text column. I  then tokenized the text and removed stop words using the tidytext package. A document-term matrix was constructed using the tm package, representing the frequency of each word in each category.

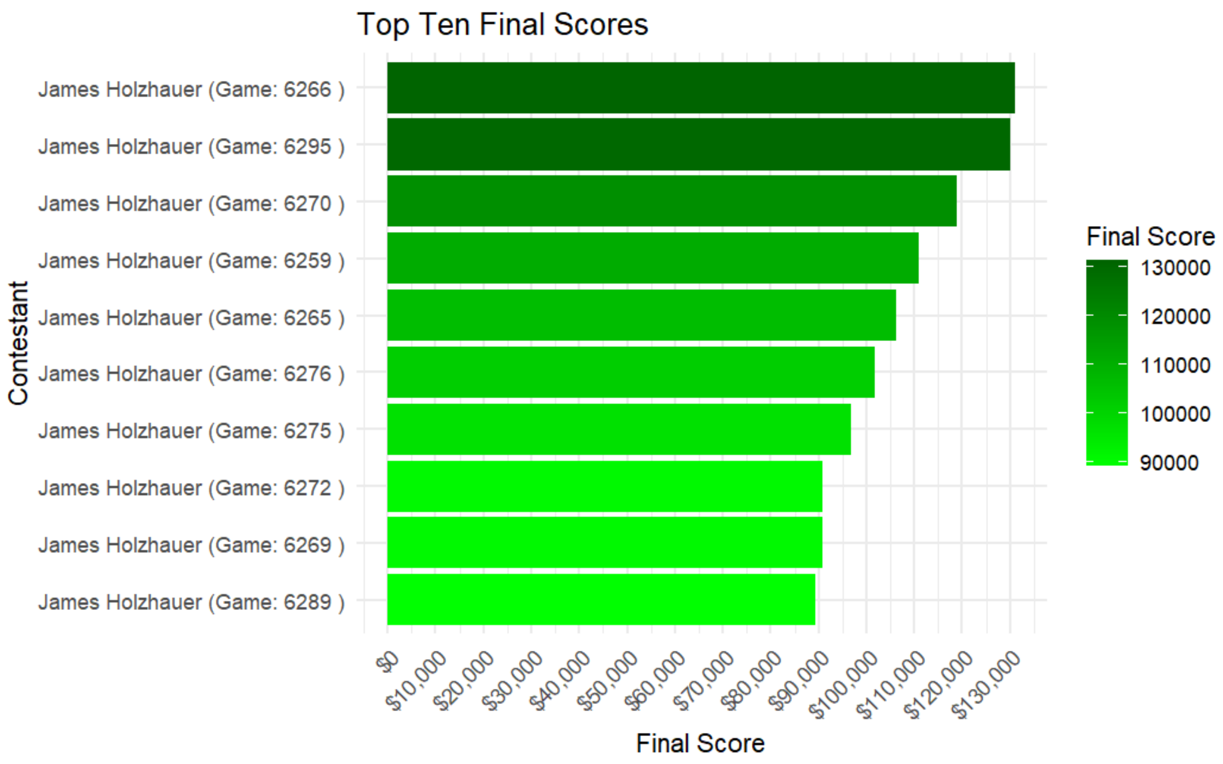
Latent Dirichlet Allocation (LDA), a probabilistic topic modeling algorithm, was applied to the document-term matrix to discover latent topics within the text data. LDA assumes that each document (in this case, a category) is a mixture of various topics, and each topic is characterized by a distribution of words. By specifying the desired number of topics (five in our analysis), LDA identified the most probable words associated with each topic. These topics were then interpreted as archetypal categories, and the top words for each topic were used to represent the semantic meaning of each category. Finally, I manually assigned descriptive names to these archetypal categories based on the top words and my domain knowledge of Jeopardy categories. The result being, for example, the category "3 Consonants, No Vowels" gets collapsed to "Wordplay" or "Largest Cities" gets collapsed to "Geography".

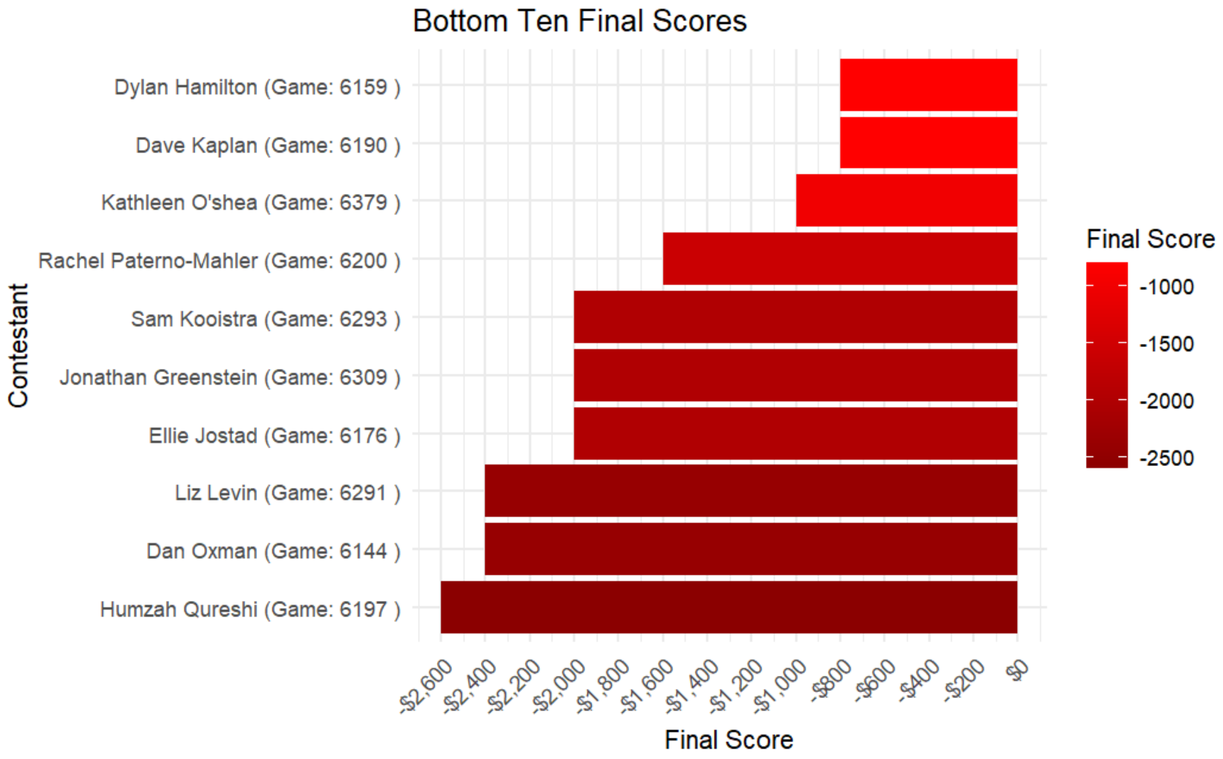
The ten gathered metacategories found were: **American History, Science and Sports, Wordplay and Other, Pop Culture, Entertainment, World History, Literature, Film, US Government, and Geography.**

## Results

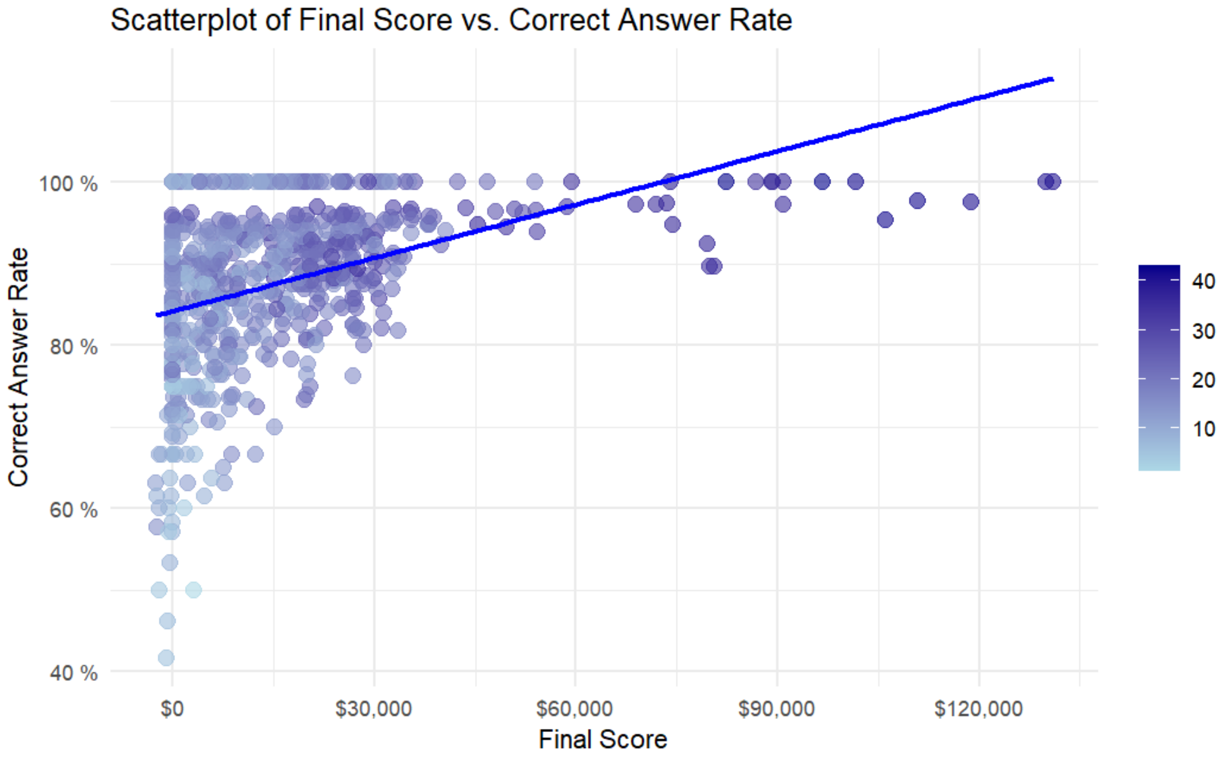


In Jeopardy!, there are special clues, called Daily Doubles, wherein the contestant is allowed to wager any amount of their current score. Additionally, in the last clue of the game, called Final Jeopardy, every contestant is given the opportunity to wager their score. This scatterplot shows the relationship between the score of a contestant before one of these Daily Doubles, and the percentage of their final score they wager. Several interesting trends can be seen in this figure. First, in Jeopardy and Double Jeopardy, several distinct bands of datapoints appear. This is due to the fact that contestants tend to wager in nearest thousand amounts, so each band corresponds to a thousand wager ($1000, $2000, etc.). Secondly, for non-Final Jeopardy wagers, as a contestant's scores increases, the percentage of their score they are willing to wager decreases reciprocally, Conversely, the percentage that contestants wager in Final Jeopardy is somewhat uniform. A potential reason for this discrepancy is in Daily Doubles, contestants often adopt a risk-averse strategy, decreasing their wager percentage as their scores increase to protect a lead. Conversely, Final Jeopardy wagers tend to be more uniform across contestants due to strategic frameworks influenced by game theory, focusing on optimizing outcomes based on the scores of all contestants. This results in more calculated, standardized bets in Final Jeopardy, where the stakes and pressure to secure a win are significantly higher.

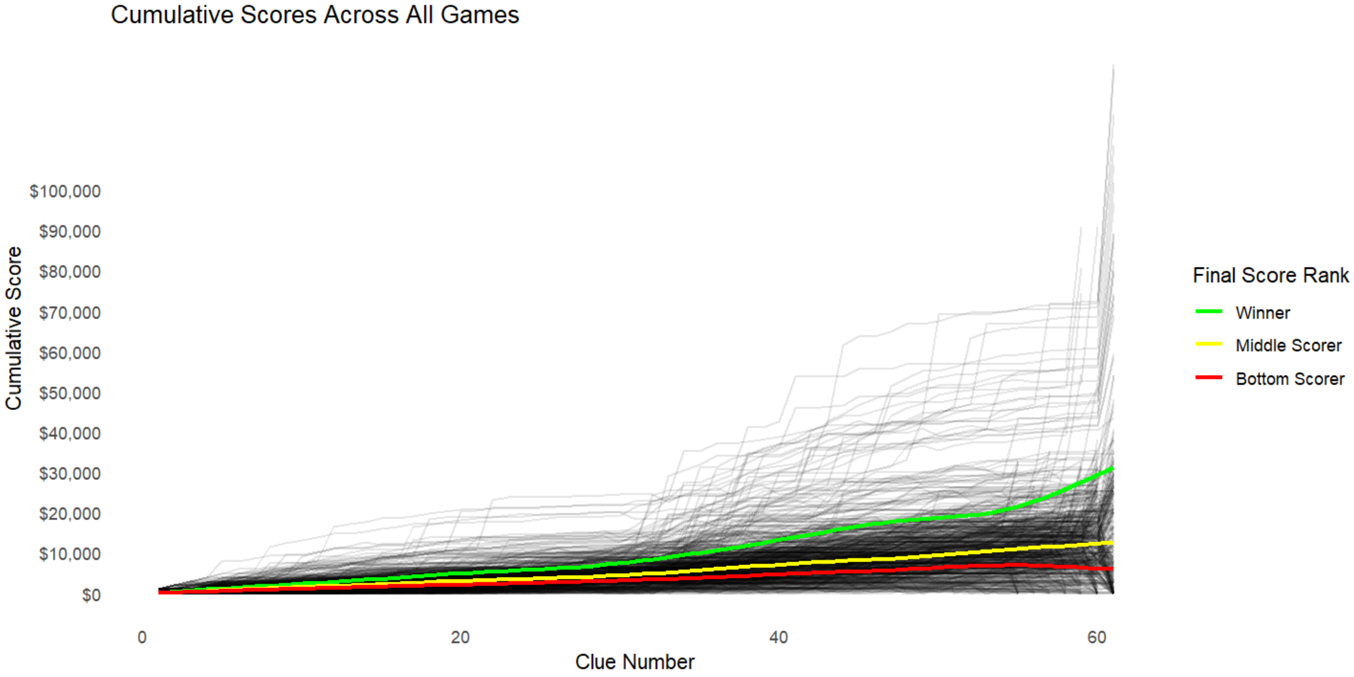




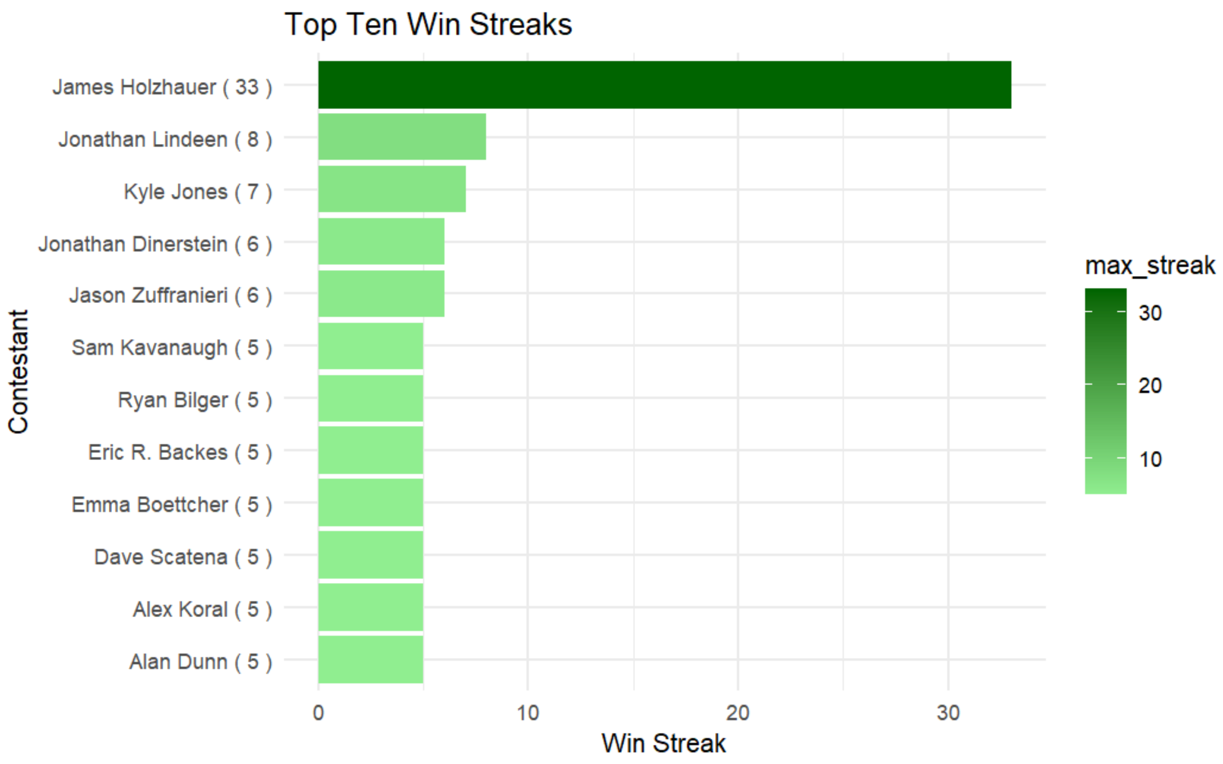
These two plots show the top and bottom 10 final scores in Jeopardy's 35th season. The top 10 scoring games are all from James Holzhauer. During this season, James Holzhauer was an unprecedented contestant insofar that he was both really really good at triva, and routinely bet all of his money whenever he got asked a Daily Double, a unique clue where a contestant would be able to bet up to their total, thus "doubling" their money. In fact, not only does Holzhauer own the top games of the 35th season, he owns the top ten winnings games of all time. All of the bottom ten scoring games in the season are negatives, and unsurprisingly, none of them come from winners.



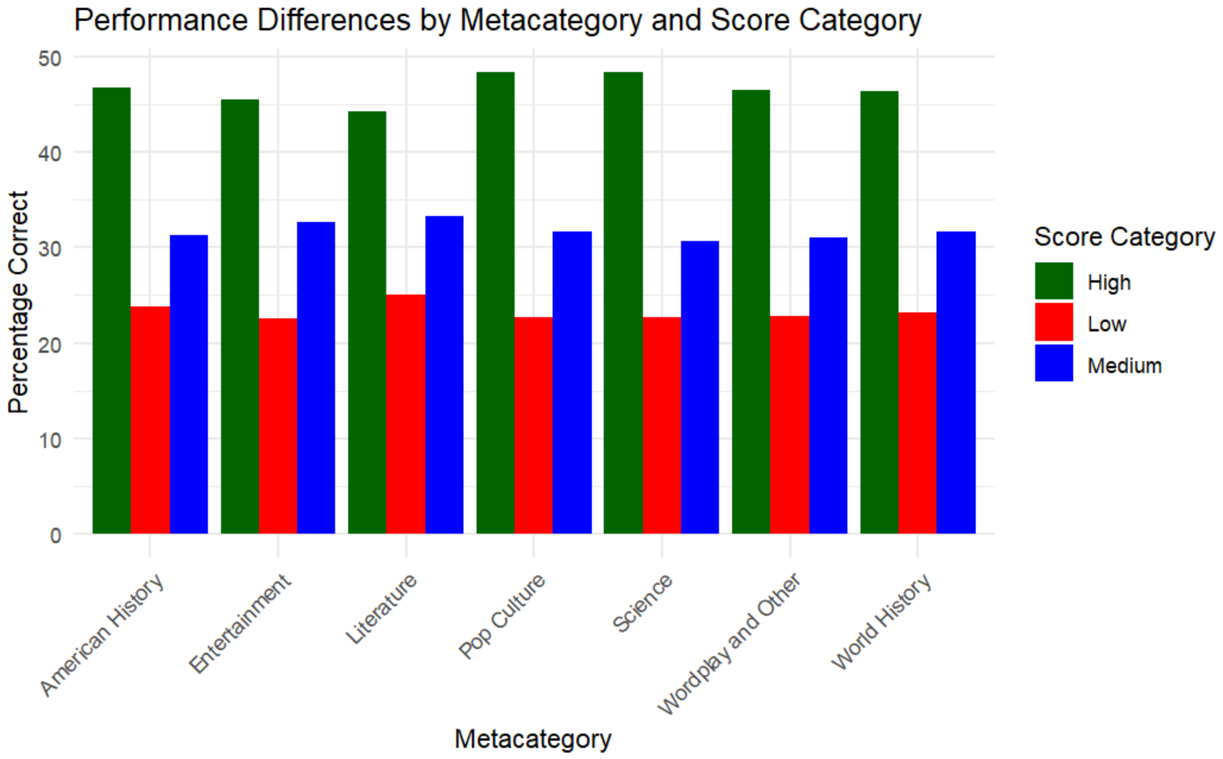
The correct answer rate represents the percentage of all questions that the contestant has answered that he has also got right. As can be seen from the plot, there is not a particularly strong correlation, only a slight positive correlation between the correct answer rate and final score. This can be interpreted in a lot of ways, but what the correct answer rate does not capture how many questions a contestant answers. Thus a contestant who answered one question in a game and got it right would have a very high answer rate, but a very low final score. However, another feature of this graph is that while for low final scores, a variety of correct answer rates can be found, once you get to higher final scores (>$30,000), you need a high (>%90) correct answer rate for this. A possible explanation for this is that every correct answer in Jeopardy deducts the value of the clue. Thus, if you want a high score you need to both be answering a lot of questions right AND not answering many questions wrong.



This figure shows the progression of every game of the season superimposed upon each other, in the black lines, and the average game for a winner, a middle scorer and a bottom scorer, colored in green, yellow and red respectively. You might notice that there are many significant (>$10,000) jumps up and down, despite the maximum value of a clue only being $2000. These are daily doubles, a special clue wherein a contestant can bet any amount of money, up to their total. At the end of the game, Final Jeopardy, everyone gets an opportunity to bet any amount of money, leading to those huge spikes for everyone at clue 61. The majority of games, where the lines are densest, end with the contestants making $0-$30,000. Those few games where the contestants win upward of $60,000 are all James Holzhauer games. The average winner won $30,000 and the average bottom scorer of a game only had a final score of $5,000. Interestingly, the winner, on average, leads the entire game, and the loser is trailing both other contestants the entire game, and also loses money on average during Final Jeopardy, as can be seen by the dip at the end.



As is already supported by the other figures, James Holzhauer is an unprecedented talen in Jeopardy, holding not only by far the longest win streak of this season of Jeopardy, but also holding the 4th highest win streak all-time. There were many people who had win streaks of 5.



The above figure compares the percentage of correct answers across different metacategories for Jeopardy contestants with varying levels of scores. The contestants are categorized into three groups: high scorers (cumulative score of 10,000 or above), medium scorers (cumulative score between 2,500 and 9,999), and low scorers (cumulative score below 2,500). Across all metacategories, as expected, high scorers have a higher correct answer rate than medium scorers, which have a higher correct answer rate than lower scorers. However, the high scorers had a higher performance in pop culture, science and history questions and a comparatively lower correct answer rate on literature and wordplay questions.

## Conclusion and Discussion

The analysis of Jeopardy's 35th season data reveals several interesting findings about the factors influencing contestants' performance and win streaks. One key observation is the relationship between Daily Double wagers and contestants' scores. As scores increase, contestants tend to wager a smaller percentage of their total score on Daily Doubles, likely adopting a risk-averse strategy to protect their lead. In contrast, Final Jeopardy wagers are more uniform across contestants, possibly due to game theory-influenced strategies aimed at optimizing outcomes based on all contestants' scores.

The data also highlights the exceptional performance of James Holzhauer, who not only holds the top 10 highest-scoring games of the 35th season but also the top 10 highest-scoring games in Jeopardy history. Holzhauer's unique strategy of consistently betting large amounts on Daily Doubles contributed to his unprecedented success. Additionally, the analysis shows that while there is only a slight positive correlation between correct answer rate and final score, high final scores (>$30,000) require a high correct answer rate (>90%). This suggests that to achieve a high score, contestants must answer many questions correctly while minimizing incorrect responses.

Furthermore, the study of metacategories reveals that high scorers consistently outperform medium and low scorers across all topic areas. This finding indicates that a broad knowledge base and the ability to quickly recall information from various fields are essential for success in Jeopardy. The analysis of win streaks also underscores James Holzhauer's exceptional performance, as he holds the longest win streak of the 35th season and the 4th highest win streak in Jeopardy history. Overall, this study provides valuable insights into the factors that contribute to success in Jeopardy, highlighting the importance of strategic wagering, a high correct answer rate, and a wide-ranging knowledge base.