Final Report

Nikki Johnson, Molly Hischke, Collin Brehmer, and Kyle Hancock

12/13/2019

## Introduction

The Daily Double is an important part of the game of Jeopardy. If a player can effectively predict where the Daily Doubles will be on the board, they have a clear advantage to win the game. Recent Jeopardy contestant James Holzhauer became known for his strategy of hopping around the board to find the Daily Doubles and then betting larger than normal quantities. In just 32 games, his total winnings reached over $2.5 million (an average of over $75,000 per game).

What positions on the Jeopardy board are most likely to be a Daily Double? Recently, several people have attempted to answer this question and their results can be found in a variety of different sources. Finding the answer to this question is not as simple as it seems at first. The Jeopardy datasets found on GitHub do not include the questions that were not asked during an episode. This can make finding the exact x and y positions difficult. Multiple approaches can be taken to deal with this issue. The approach taken to develop the figures in this report are described in the methods section. The answer to this question can be visualized through a heatmap of the Jeopardy board that indicate where the most and least daily doubles were located. Similar to previous analyses, the results in this report are presented in a heatmap to determine if there was a difference in daily double position by round. As a secondary question, this analysis faceted the results by year to evaluate changes over time, which has not been included in previous analyses. This important question presents a challenge in data cleaning, but when done correctly can result in interesting figures that have drawn the attention of many Jeopardy fans and data enthusiasts.

## Methods:

The Jeopardy dataset utilized was found on GitHub which included the air date, round, daily double information, category, value, question, answer, comments, and notes. The dataset was filtered to only include dates beyond November 26, 2001. On that date, there was a change in the point values within the rounds (for example, round one went from values of 100, 200, 300, 400, and 500 to 200, 400, 600, 800 and 1000). During the point value update, there could have been other changes made to Jeopardy impacting the daily double locations (i.e. changes in how the positions were assigned). Therefore, we have excluded dates prior to the change.

There were two main challenges in finding the positions of the daily double questions:

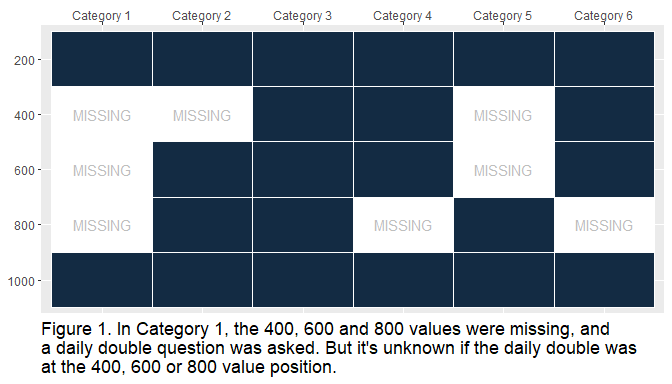
1. Finding the x position (the location of the categories).
2. Finding the y position (the location of the values).

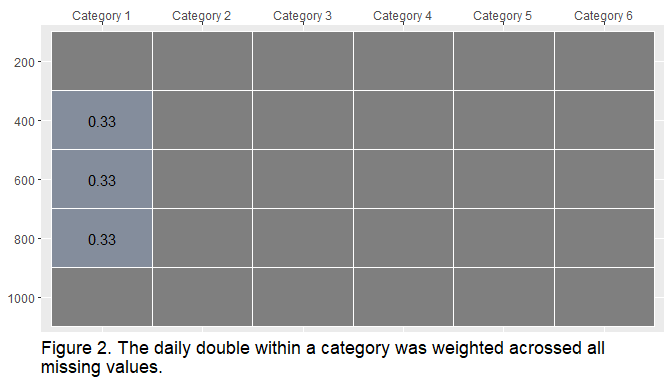
Not all categories (n = 12) were available in the dataset for all episodes. To solve this issue, only episodes where all 12 categories were played were included. The assumption was that the dataset listed the categories in order by x location within episode (i.e. first category to appear was assigned a x position of 1).

In addition, not every question (n = 5) within a category was available in the dataset. This was especially challenging when a category contained a daily double question. Since players select their own bid for daily doubles, the value within the dataset gave little information about the location on the y axis. If there were 2 missing questions and 1 daily double within a category, the daily double could be at any of the missing value locations (Figure 1). To solve the issue, the daily double was weighted across the unasked questions in the category (i.e. 2 missing values in a daily double category would result in a 1/2 daily double at each missing position).

The Jeopardy Round 2 has two locations with daily doubles. We did not check that there were two daily doubles asked within round 2; therefore, we did not weight for a missing daily double value across the board.

All of the code for this analysis can be found in the GitHub repository found [here](https://github.com/cbrehmer97/ERHS535_final_group5).The Flex Dashboard.Rmd produces the interactive heatmaps by year and round. The code that cleans the data into a tidy format is in [cleaning\_code.R](https://github.com/cbrehmer97/ERHS535_final_group5/blob/ERHS535_final_group5/cleaning_code.R). The code that produces the heatmaps by year and round is in [figure\_code.R](https://github.com/cbrehmer97/ERHS535_final_group5/blob/ERHS535_final_group5/figure_code.R).





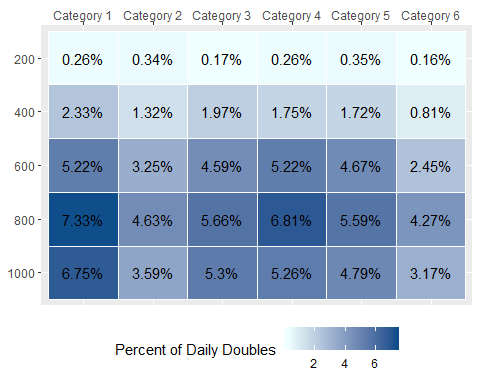
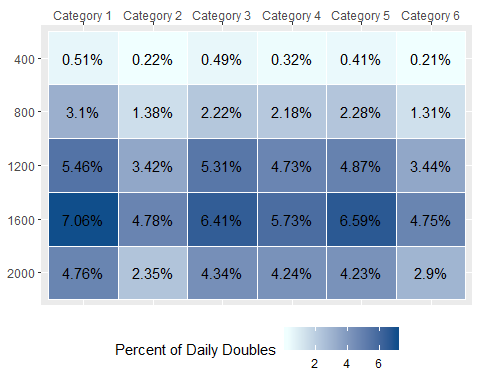
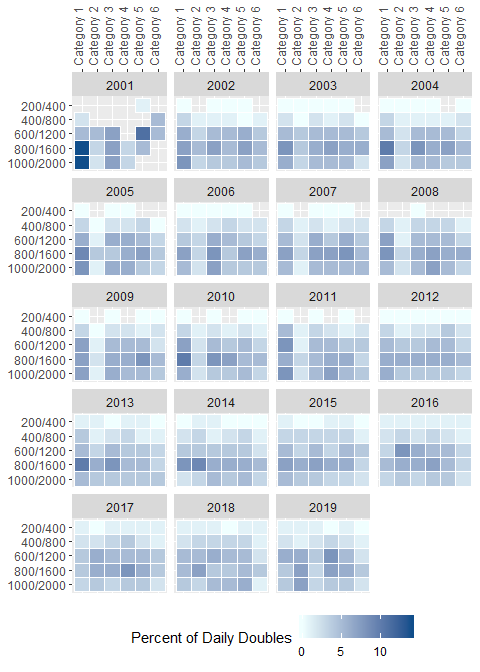
## Results

The heatmaps displaying the percent of daily doubles that occurred at each position is given in Figure 3 (round 1) and Figure 4 (round 2). In round 1, the daily double occurred most often (7.33%) in category 1 at a point value of 800. Other points where the daily double occurred frequently (> 6%) were category 4 at a point value of 800 (6.81%), and category 1 at a point value of 1000 (6.75%). The least likely placements were across all categories in value 100 (0.35% or less), as well as in categories 2 (4.63% or less) and 6 (4.27% or less). Value 800 is favored across all categories (range: 4.27% – 7.33%).

In round 2, the daily double was located most often in category 1 at a point value of 1600 (7.06%). The other common positions were category 5, point value 1600 (6.59%) and category 3, point value 1600 (6.41%). The least likely occurrences were across value 100 (0.51% or less) and down categories 2 and 6 (< 4.78% and < 4.75% respectively). The daily double appeared most frequently at the 1600 point value across all categories (4.75% – 7.06%). Overall, the occurrence of the daily double remained extremely similar between round 1 and round 2 for the years analyzed.

As a secondary question, the heatmaps for all data (rounds 1 and 2 combined) were faceted by year to look at changes over time (Figure 5). Because the locations of the daily doubles were similar between rounds, the data was not separated by round for the year analysis. The results for years 2001 - 2015 look similar to the results by round. The “hot” points are similar as well as the overall pattern. In 2016, there seems to be a shift away from the traditional spots. Previously, category 2 was one of the least likely columns but after 2016 the percentage surpassed the percentages in category 1 (previously, the most likely column). There also seems to be a shift from row 4 to row 3. Before 2016, rows 4 and 5 were the most likely to contain a Daily Double. After 2016, row 3 had a higher percentage of Daily Doubles. The pattern after 2016 also seems to have more randomness than prior to 2016. In several of the heatmaps there are missing tiles. The year 2001 has the most missing tiles because there was only one month of data from this year. In the other years with missing tiles, they are all in the first row. Based on previous analyses and our results, it is likely that these missing tiles are there because a Daily Double was never found in that position. There are likely more positions that should be missing but weighting caused there to be a very small percentage in that spot. The analysis by Nathan Yau (referenced later in the conclusions) found that there was a maximum of 5 Daily Doubles in any of the spots in the first row over the course of 31 seasons.

The heatmap results could underrepresent some of the daily double locations because (1) we did not check that two daily double questions were asked in round 2 or include a weighting to account for this missing question, and (2) we filtered to include the episodes were 12 categories were asked (excluding a subset of data points). The unaccounted for daily doubles in both cases could have influenced the gradient of the heatmaps for round 2 (subsequently impacting trend results in the year analysis).

  
Figure 3. Percent of Daily Doubles by Position in Round 1  
  
  
  
  
Figure 4. Percent of Daily Doubles by Position in Round 2  
  
  
  
  
Figure 5. Percent of Daily Doubles by Position Faceted by Year

## Conclusions

Our findings are very similar to previous analyses on Daily Double location by round. One of the original analyses of Daily Double location was conducted by Nathan Yau (<https://flowingdata.com/2015/03/03/where-to-find-jeopardy-daily-doubles/>) in 2015. The results of this study look similar to his heatmap, which shows the location of all Daily Doubles for 31 seasons. Yau’s analysis was done because of a Jeopardy contestant, Arthur Chu, that appeared on the show in 2014. Arthur Chu found success on the show by jumping around the board hunting for the Daily Doubles. He was not the first to adopt this strategy, but was one of the most controversial players because of his approach. Since 2015, there have been other analyses that looked at the same question and found similar results by round. To our knowledge, no other analyses have included heatmaps or data by year. This analysis included data up to 2019 and showed a noticeable shift in location around 2015/2016. This shift could have occurred because of Arthur Chu and the subsequent analyses of Daily Double locations. Future analyses may consider evaluating this change over the next few years to see if another pattern is established, or if Jeopardy adopts a more random Daily Double placement.