Jeopardy: Most Repeating and Least Repeating Info

By Kenny Nguyen

I. Introduction

Jeopardy has been a long time running game show in America. There has a been multiple questions that have been brought up during the show and they each have multiple elements that define them. These elements can be defined multiple ways but we'll define them in questions format to determine what information can be obtained from them. One question that can be asked is, "What are the most recurring recurring answers on the show?". Another is ,"What categories do these questions mostly go to?" "What is the most used money value question on the show?" "What are the highest money values question on the show?". We'll also ask the reverse of all these questions to see the least occurrences for each one and some of the lowest money value questions on the show. All of these can be answered with a dataset that spans from 1984-09-10 to 2012-01-27.

II. Motivation and background

Jeopardy, as said before, is a long running show in America and it's interesting to see how much answers are repeated and how many different money values they can use. We can also see how many tiebreakers total have ever shown up in the show since they barely show up and are extremely rare.. There's multiple things to consider when looking over a gameshow this long and it is filled with information that is ready to look at.

III. Data set

The dataset I use is from a user on reddit that goes by the name of trexmatt. Credit goes to him or her who profile is this link. https://www.reddit.com/user/trexmatt

He made a post on with a json file that he obtained by crawling through the website of http://www.j-archive.com/ around four years ago. His reddit post is the following link: https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/?st=jgxfzw8n&sh=b83c28fd

All credit goes to trexmatt for crawling through the website and making a json of all the information that allowed me to analyze.

IV. Methodology

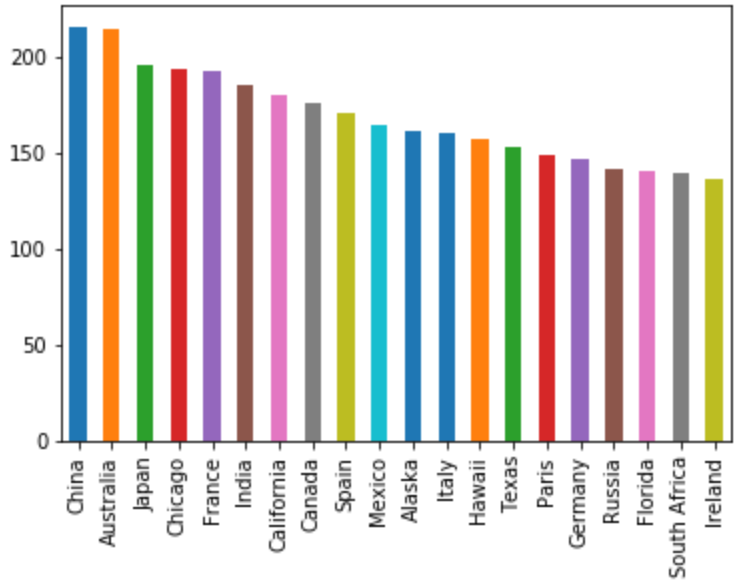
1. Import json, numpy, matplotlib, and pandas.
2. Import other necessary files if needed
3. Read the json file into the dataframe
4. After reading the json file, we can look at the dataframes to see all the columns with information
5. Next we can look at how far this data set spans by sorting the values with sort_value(by='air_date'). This sorts our dataset by the date they were air and we can see the beginning and the end at the same time.
6. Looking at the rounds columns, we can use value_counts() to see how many times tiebreakers, regular jeopardy, double jeopardy, and final jeopardy showed up.

7. Doing the same as step 6, we can look at all the categories that showed up. However, it has a greater amount of data in it so our dataset won't show everything but some of the top occurring ones and some of the least occurring ones. We are still able to get tons of information from this.
8. Repeating step 6, we do this with the answer columns to see most occurring and least occuring answers.
9. We'll make a different dataset with only the values column which we'll separate by data[data.columns[6:]].replace('[\$,]', '', regex=True).astype(float). This allows us to replace the $ with blanks to let us sort the money values easier.
10. Knowing that some of the money values are NaN, we would do dropna on any rows that has it and put inplace = true to make sure our second dataset has all the information needed to look at our money values. We would then sort the second dataset by values and we can see the lowest values used and the highest values used.
11. Lastly we can make a bar graph for each one to show our dataset in a clearer image. We would the do the categories and answer column with data['Insert column value'].value_counts().head(20).plot.bar() because we would simplify those datas to just the top 20. Then instead of .head(20), we would do tail(20) for the least occurring ones. For rounds, we wouldn't need to the 20 in head and tails because there are only 4 rounds so we can just do .head() and .tail() instead. For highest values ever, we would use our second dataset and call it with sort_values(by=['value']).tail(20).plot.bar() because it needs to be sorted and it will plot the values on the y axis and which question it is on the

x axis. These would be the highest values and to do lowest, we would do head(20) instead.

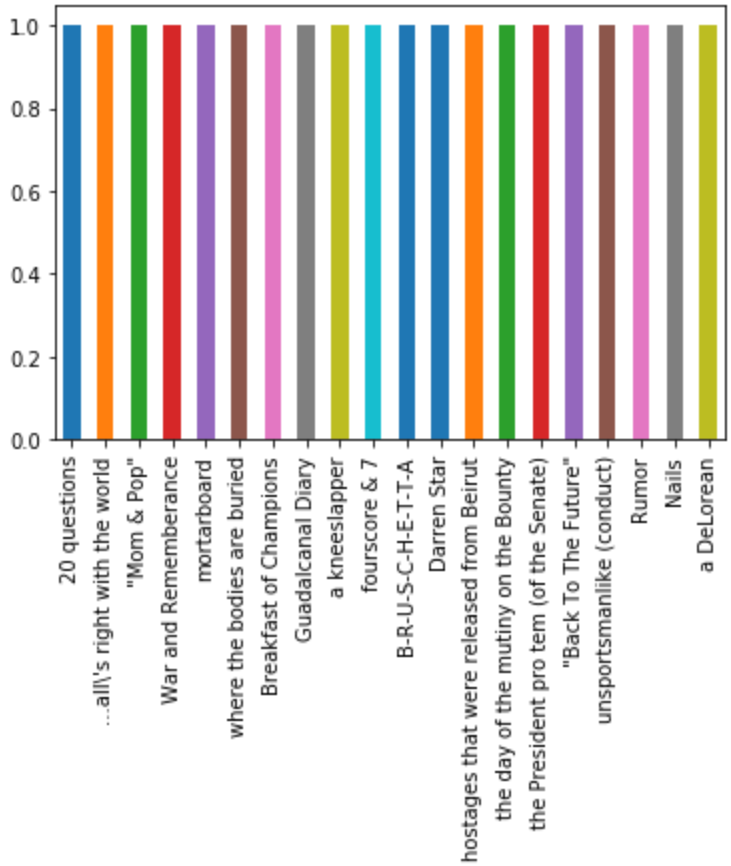
V. Results

A. Top 20 Answers



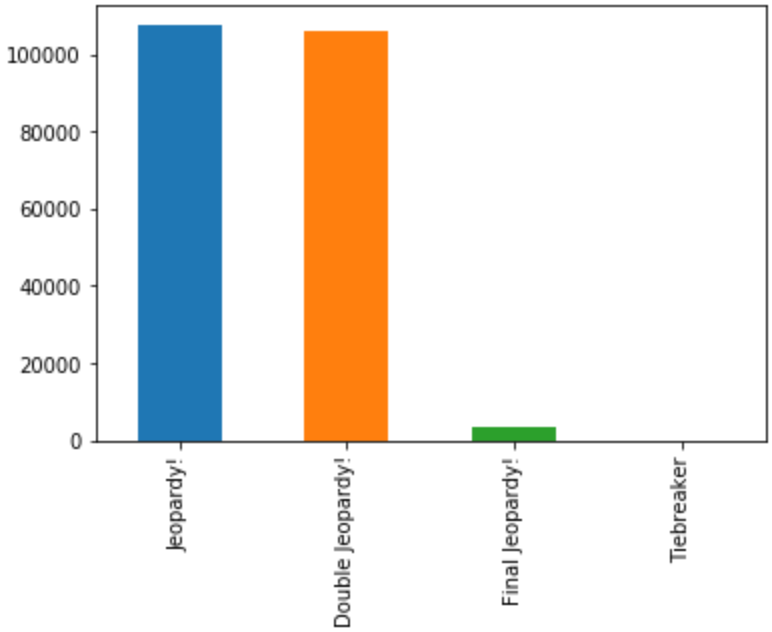
These are the top 20 answers and we can see that the most occurring ones are like states and countries names because they were most likely used for specific questions such as “What place has the most population” etc. They are the most generic and can be used for multiple questions.

B. Least 20 answers



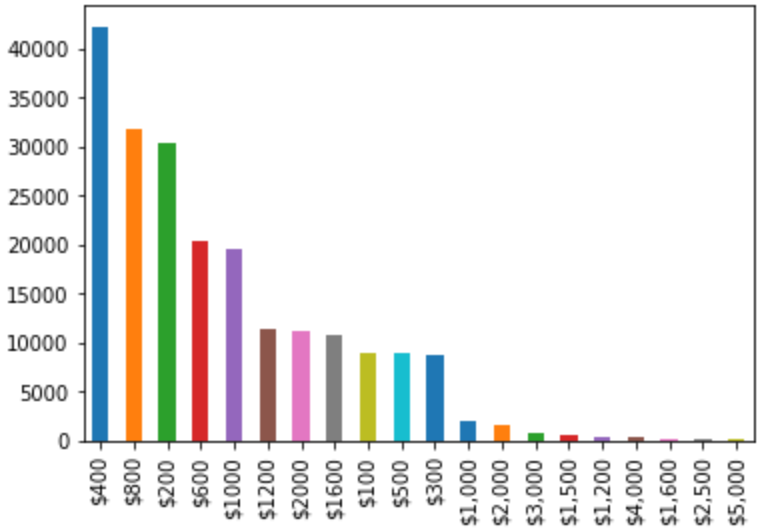
These are some of the least occurring question sand we can see that they are oddly specific, part of a sentence, or something that no one most likely knows.

C. Rounds



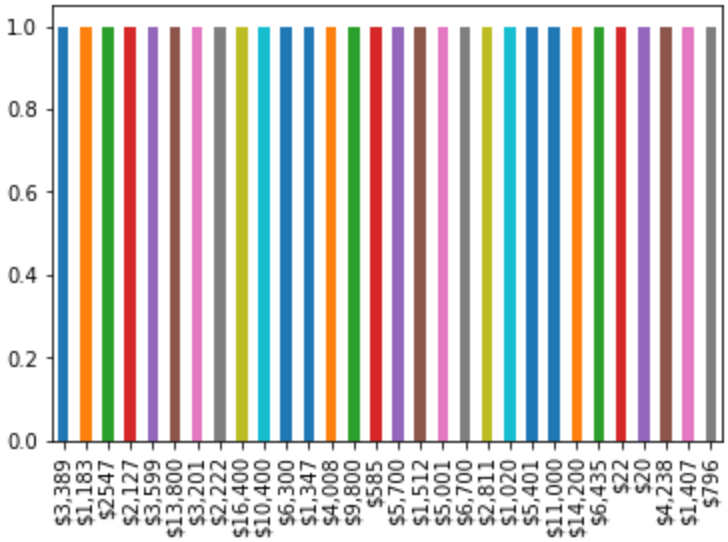
For the jeopardy rounds, we can see there are numerous Jeopardy and Double Jeopardy. Mostly because the show is built upon those questions and final jeopardy is considerable less because it only shows up near the end. While tiebreaker, when looking at our data, is 3 because the chances of two people having the same amount of money near the end is near impossible. We can see in our data graph that it is near impossible to see and we would assume it would be 0.

D. Top 20 occurring values



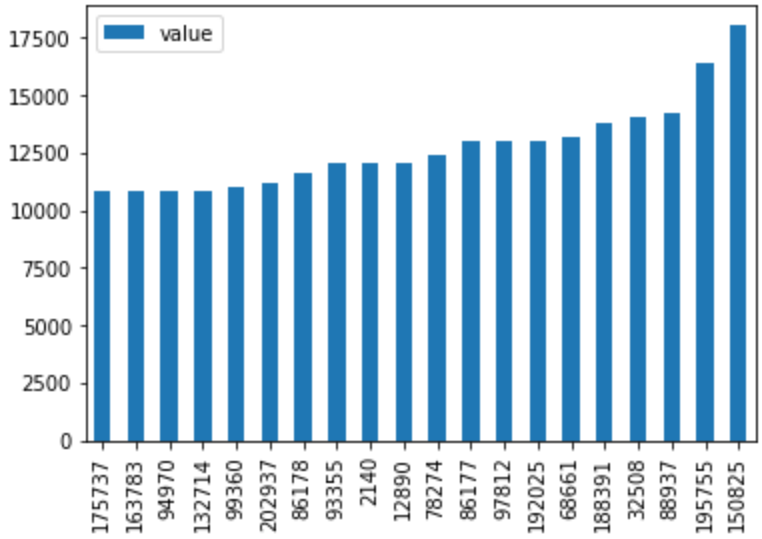
In this graph, we see that the most occuring value is $400. Then we can follow the graph and see the rest and the 20th spot being $5000. The $400 value is the most occurring one because the original rounds of jeopardy were form $100-$500 then they were doubled to $200-$1000 on November 26, 2001. Then from there, $800 most likely barely beat out $200 because of double jeopardy rounds and daily double.

E. Some of the 20 least occuring values



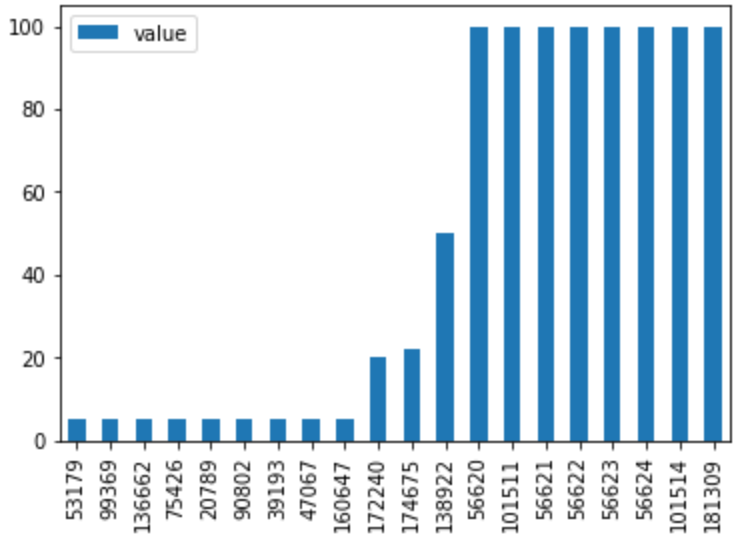
Some of the least occurring values in jeopardy are the following we can see on the graph. They are mostly randomly numbers that occur from daily double where the contestant can wager any amount they want out of the amount they have.

F. Top 20 Values



These are the top 20 money values that has shown up on the show. The values are on the y-axis which is on the left. While the bottom shows which question number that value was on. It goes from roughly ~ 10000 to 18000 for the top 20 money values. These were most likely all used for daily double because the amounts are absurdly high.

G. Some of the 20 least values



These are some of the lowest amount of value earned from a question and they are just basic jeopardy questions. We can see it ranges from like 5 to 100.