This.. is.. Jeopardy

Team

- Anand Rajagopal
- Anubhav Gupta
- Joshua Appleman
- Juan Shishido

Project Goals

- Learn about gender inequality in Jeopardy! its causes and how it is changing throughout the seasons
- Discover the characteristics of the game's best players
- Understand how the likelihood of winning Jeopardy! changes as the game progresses
- Explore the difference episodes throughout the seasons and how dramatically the game dynamics vary
- Understand the difference in risky behavior between contestants in 1st, 2nd and 3rd place and how that has changed throughout the seasons
- Learn fun and interesting facts about the game show and its history

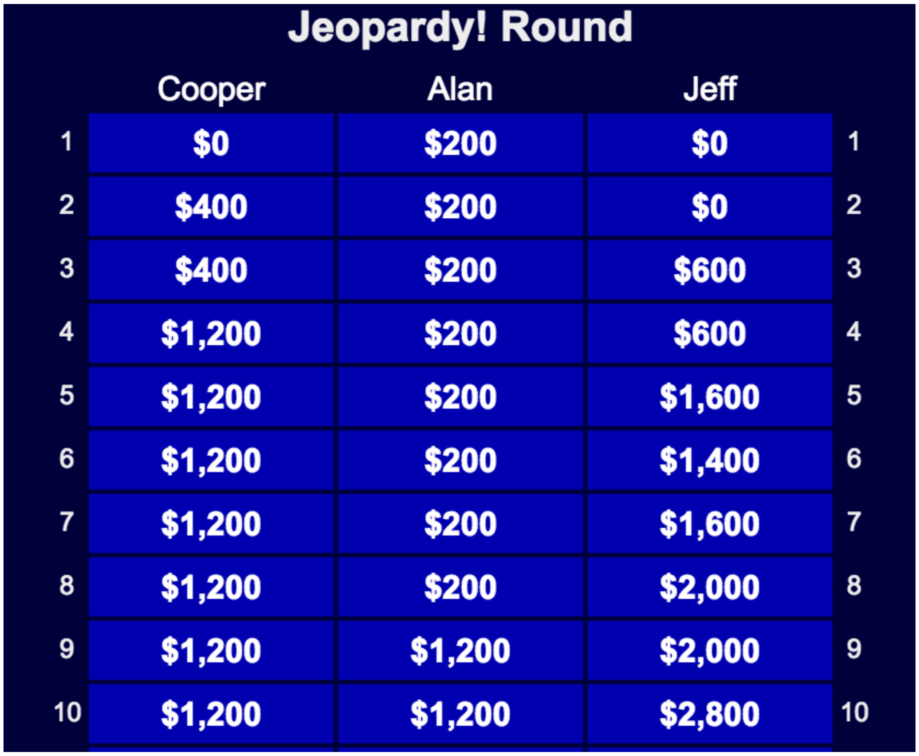
Data

The initial data set, found on Reddit [1], included qualitative data on over 200,000 Jeopardy! questions. It had data on the round that the question corresponded to, the questions and answers themselves, the dollar value of the question, and the show number and air date.

In our exploratory data analysis phase, we found several interesting results. For example, we found that the most frequent type of answers were related to geography. We also noticed, while looking at the average question value across years, that there was a large increase between 2001 and 2002. In 2001, the average question value was $496. It increased to $940 in 2002. With additional research, we found that the values doubled on November 26, 2001. This would inform some of the subsequent decisions we would make with respect to comparing episodes across time.

In processing the data, we also found that the number of episodes varied by season. This was not a function of the show, but a result of data not being available in the earlier years. The source of the data set posted on Reddit was from J! Archive, a "fan-created archive of Jeopardy! games and players."

Because Jeopardy! is not just about the questions and answers, we decided to obtain additional data to complement what we already had. From J! Archive, we scraped the individual episode scores. An example of the data is shown below.



Jeopardy! Round

	Cooper	Alan	Jeff	
1	$0	$200	$0	1
2	$400	$200	$0	2
3	$400	$200	$600	3
4	$1,200	$200	$600	4
5	$1,200	$200	$1,600	5
6	$1,200	$200	$1,400	6
7	$1,200	$200	$1,600	7
8	$1,200	$200	$2,000	8
9	$1,200	$1,200	$2,000	9
10	$1,200	$1,200	$2,800	10

This table shows the scores for each contestant on the first 10 questions in the Jeopardy! round for a particular episode. For each episode, we collected the scores data for every question in each of the three rounds. Not only did this provide the question-by-question scores as well as the total earnings, it also gave us a chance to explore the wagering dynamics of the Final Jeopardy! round. This quantitative data was used in the four of our visualizations: the heat map, the game line plot, the scatter plot, and the bipartite graph.

We also scraped the top 50 contestants and their earnings from the hall of fame section on jeopardy.com, "http://www.jeopardy.com/showguide/halloffame/50kplus/"

Tools

In order to process the data, we used both Python and Excel. The exploratory work had two components. We used IPython notebooks, reading the data into pandas DataFrames, to both transform the data and extract features we would like to use in the visualizations. We used Tableau to do deeper exploratory analysis of the data. We also used IPython notebooks for testing code for scraping the scores data. For this, we used both the pandas and BeautifulSoup modules. When we were done testing, we created Python scripts that were run on Harbinger.

For the visualization, we used: HTML/CSS, JavaScript, D3, HighCharts, JustInMind, Rhinoceros (CAD vector software good for tracing images to bring into Illustrator), Photoshop, and Illustrator.

Process and Related Work

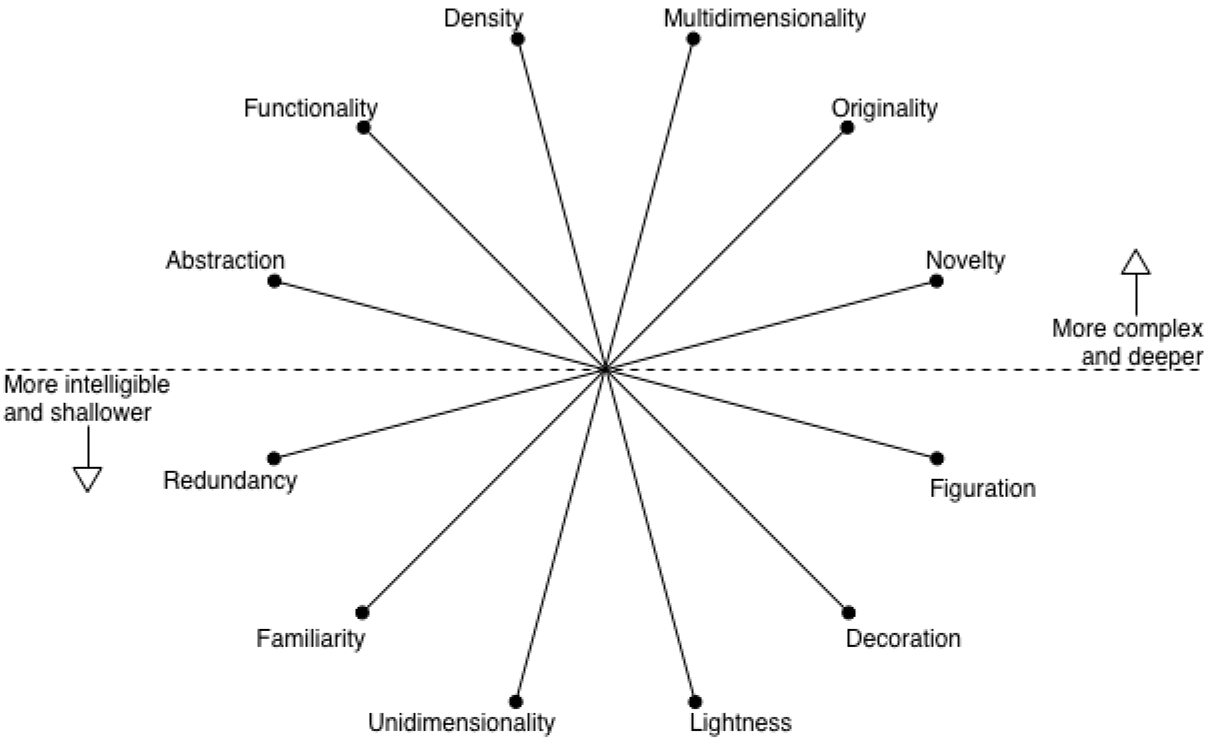
Assigning Gender to Contestants

In our data set we had 9101 unique contestants but we could not tell their gender from our web scraping. Of those, there were 1930 unique first names. To approximate their gender, we used their first names. First we downloaded a database from http://www.ssa.gov/oact/babynames/limits.html. That data set has the 1000 most popular new born baby names in the USA each year going back to 1880. It covers 74% of the US population. It has the name, sex and number of births. Some names obviously appear in both genders, such as Jordan for example, but in those cases we just assigned gender based on what was more probable.

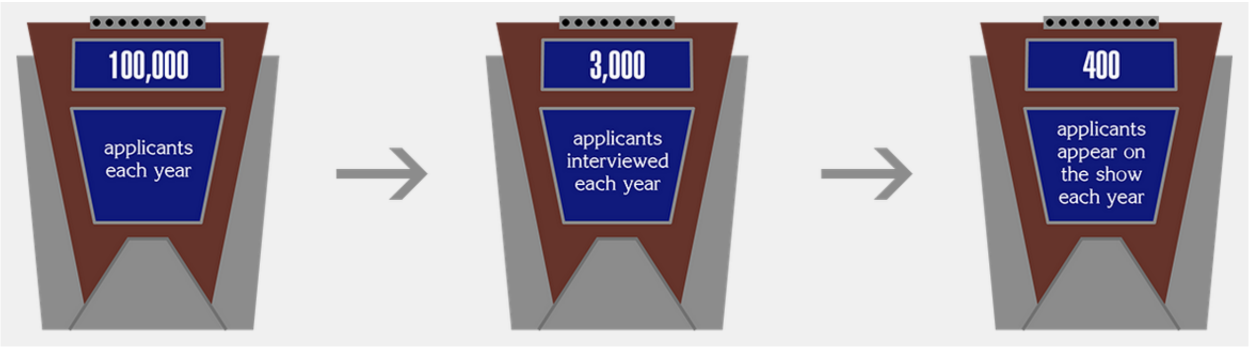
By the time this process was finished, we still had about 500 names left that did not show up in the database. The next step was using websites and APIs like http://genderchecker.com, https://gender-api.com, and https://genderize.io. The free versions of these had limitations with how many names one could enter at a time and how many could be checked in a day. Running the names through this process got us down to 100 unknown names. At this point we manually typed the names in LinkedIn and Facebook and estimated if there were more female or male results based on the pictures. Some contestants have their picture on http://j-archive.com and we manually used that as well. With this new information, we were able to give all contestants a gender.

Contestant and Gender Infographics and Line Chart

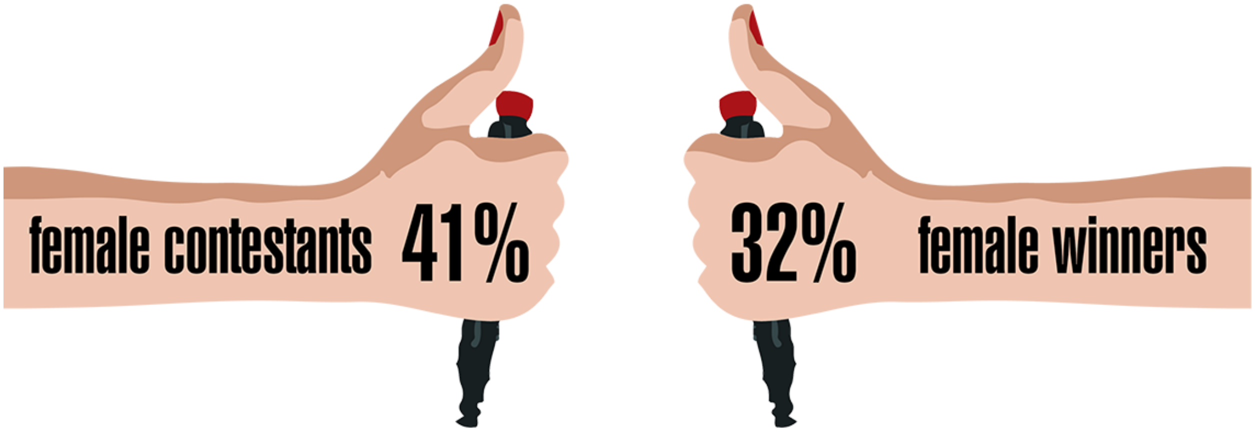
We did not want to immediately present the people visiting our visualization with something too dense and complex. We have deeper more multidimensional visualizations later on but we wanted to ease people into it. Narrative was important to us so we wanted to start lower on Cairo's (p.51) visualization wheel [1] then move upwards.



The narrative begins by telling people a bit how people end up getting on the show. We found pictures of the Jeopardy podiums on Google Images, traced them in Rhinoceros (a CAD software) and imported the vectors into Illustrator to add color and text. The podiums were a way to start out the narrative in a playful way with simple numbers. The illustrations are familiar, light and decorative. The numbers provided are unidimensional. The fonts displayed are the actual fonts used on Jeopardy.

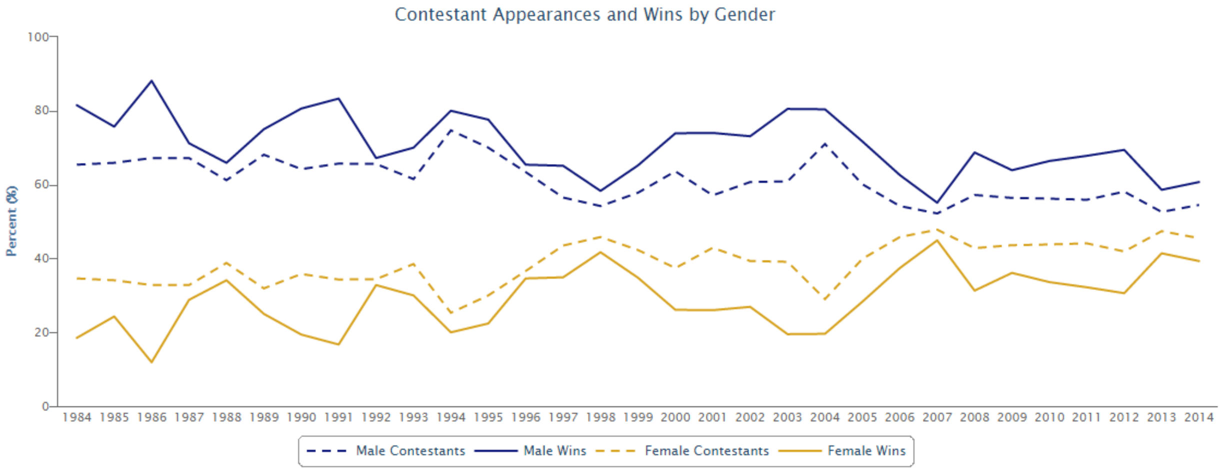

100,000
applicants each year
3,000
applicants interviewed each year
400
applicants appear on the show each year

We then quickly transition into the ‘so what?' part of the narrative. In the illustration of the hands holding buzzers, we are pointing out that not only do fewer females win than males, they win less in proportion to the number of female contestants. Nail polish matching the buzzers was added to the thumbs to show femininity.

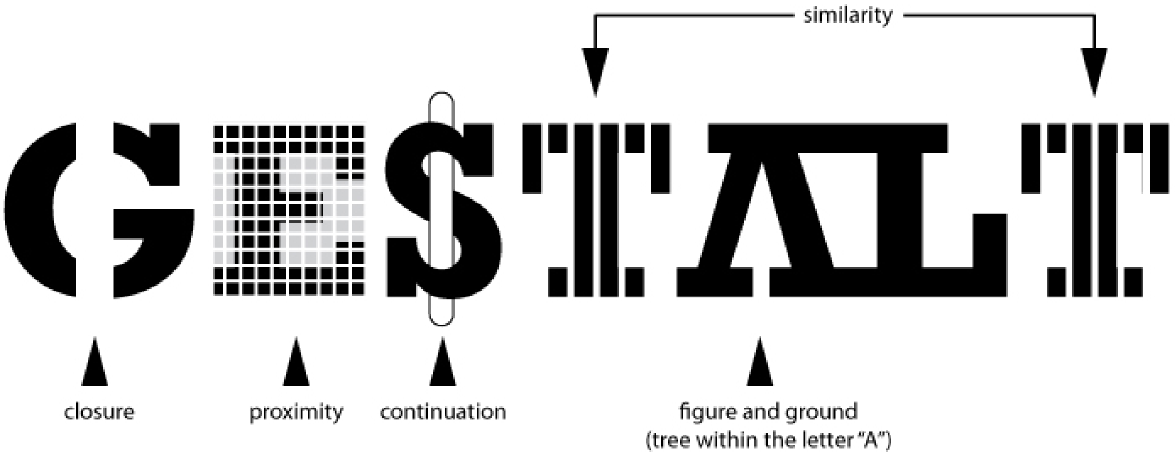


The first chart we have on the page shows the percent of female contestants and the percent of female winners from 1984-2014. Looking at Few's chapter on time-series analysis (p.146) we see that the height of the chart is important. Making it too short will make the variability difficult to detect. Making it too tall can exaggerate the variation and mislead the readers. Our goal with the height was to balance in between those extremes to show more females are coming onto the show and winning but the changes are not drastic. We are also using the Gestalt principle of similarity. The yellow color for female and blue for male matches the human isotypes in the next graphic. And the line type for contestant percentage and win percentage matches across genders.

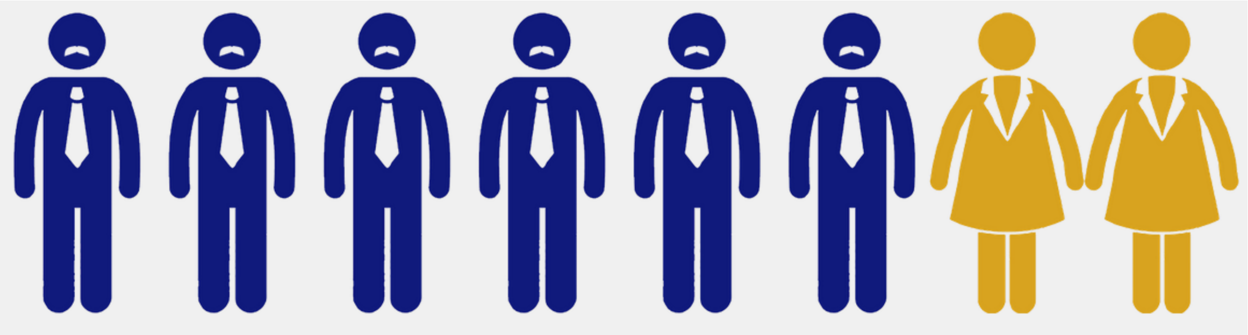
Another project from Few that was helpful in making this visualization was the stacked area chart on government spending (p. 305) [2]. In particular the suggestion to use tooltips to give extra information without cluttering the whole graphic was helpful.



Contestant Appearances and Wins by Gender



Inspiration for the isotypes of people came from the project Cairo cited by Otto and Marie Meurath about Home and Factory Weaving in England (p.72). [3] The goal is to communicate a simple idea with clarity and power. Showing the stark contrast in the number of male and female writers should send a strong message about what could be causing the gender gap in Jeopardy! wins.

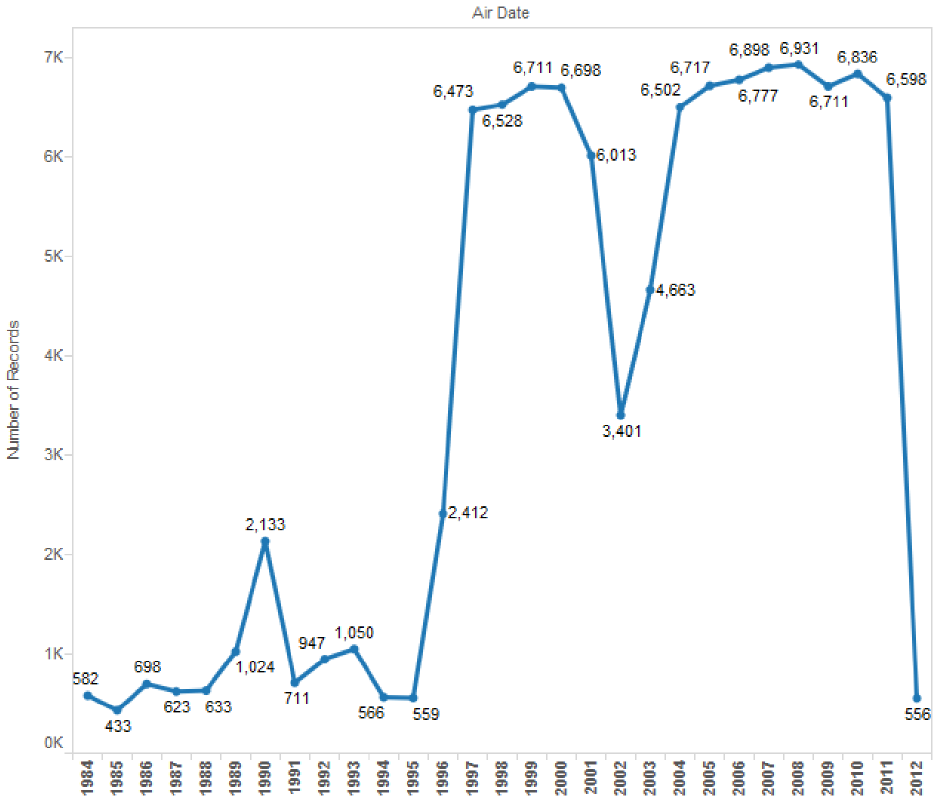


Exploratory Data Analysis using Tableau

The jeopardy dataset from J Archive was pretty vast and had information covering different aspects of the game show. Also given that, arguably, the main component of the data - the questions and answers; was text and therefore much more challenging to visualize. However we wanted to investigate if there were any underlying trends to this data. An overlying question or hypothesis that we were trying to answer was: **Is there a smart way to study/prepare for Jeopardy?**

Step 1: A logical point was to look at the distribution of the data.

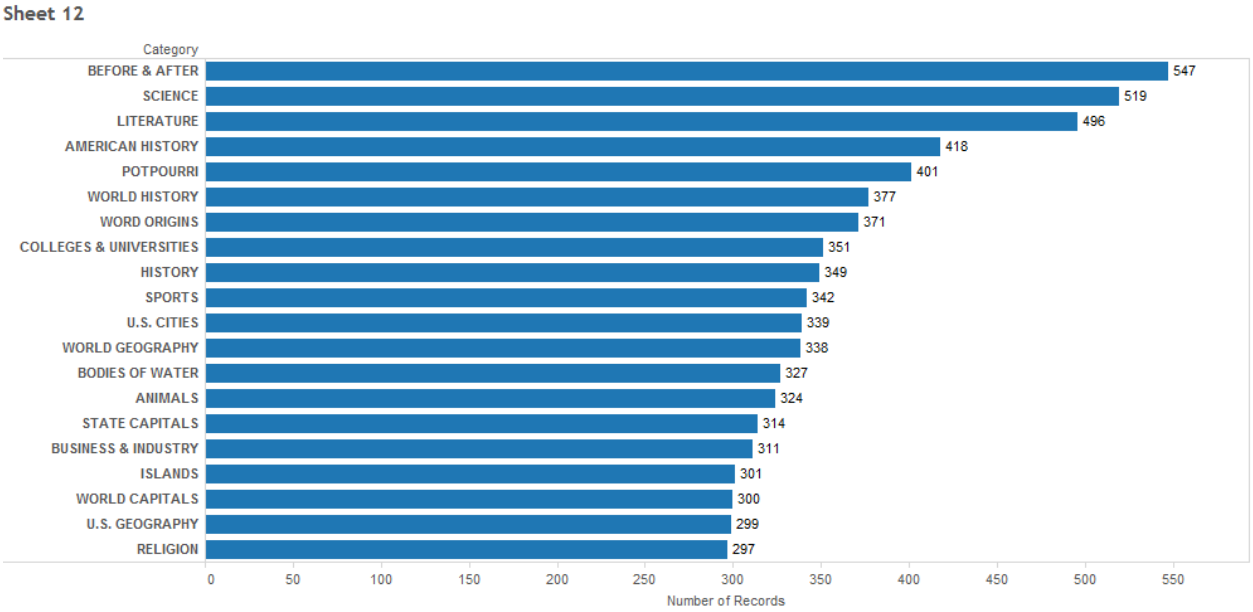
No of records by Airdate



The trend of sum of Number of Records for Air Date Year. The marks are labeled by sum of Number of Records.
The data is filtered on Round, which keeps Jeopardy!.

This helped clearly see that there was a huge skew in the data. This data has been mostly created by voluntary work of fans and this has been more diligently done after 1995. To deal with this, we decided to analyze all our data by grouping them into 3 categories - 1984-1995, 1996-2002 and 2002-2012. The second dip was probably created as an aftermath of the 9/11 attacks when there was either a dip in the airing of the shows or in its recording and archiving by fans. This would help us identify more meaningful trends in the data and rule out any bias created by the lack of data.

Step 2: Next step was to look at the categories over time to try and see if there any categories that are larger in proportion so that they can be considered more prominent. Traditionally fans have always suspected "Potpourri" to be a very popular category, but our analysis showed it different.

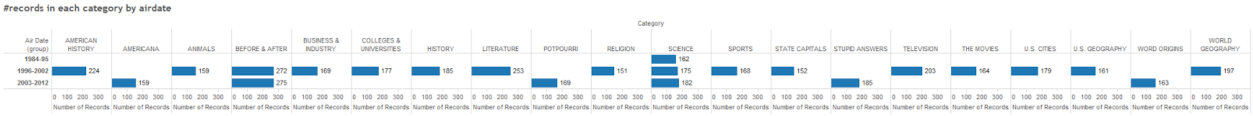


Sum of Number of Records for each Category. The marks are labeled by sum of Number of Records. The data is filtered on Round, which keeps Double Jeopardy!, Final Jeopardy!, Jeopardy! and Tiebreaker. The view is filtered on Category, which keeps 20 of 27,995 members.

This shows "Before and After" to be at the top followed by "Science", "Literature" and "American History" before the crowd favourite "Potpourri".

Step 3: To give this more context, we plotted how this varies by season as well. This provides some interesting insights. "Before and After"did not exist as a category until 1996. Yet, after that it has gone on to become the most popular category. In almost all the cases, there seems to be a peak in 1996-2002 even though this is not the largest category in terms of duration. This seems to have been a relatively less imaginative period in the history of the show where categories were repeated more often.

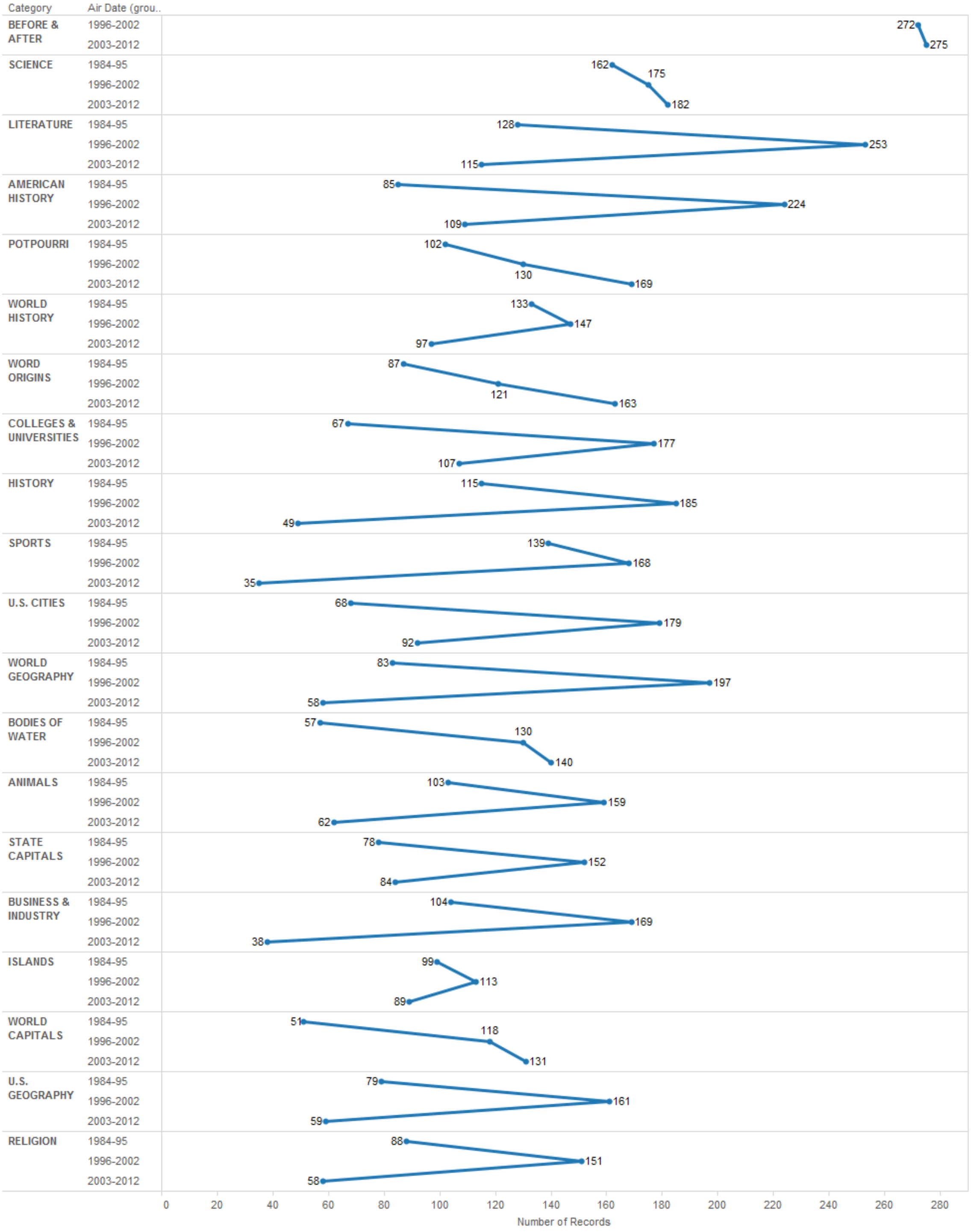
Another insight is how many topics seem to have an overlap. Many are vague, others overlap, and most seem to relate to Geography. Is there something to that?



Sum of Number of Records for each Air Date (group) broken down by Category. The marks are labeled by sum of Number of Records. The data is filtered on Round, which keeps Double Jeopardy!, Final Jeopardy!, Jeopardy! and Tiebreaker. The view is filtered on sum of Number of Records, which ranges from 150 to 275.

Another view of the same data:

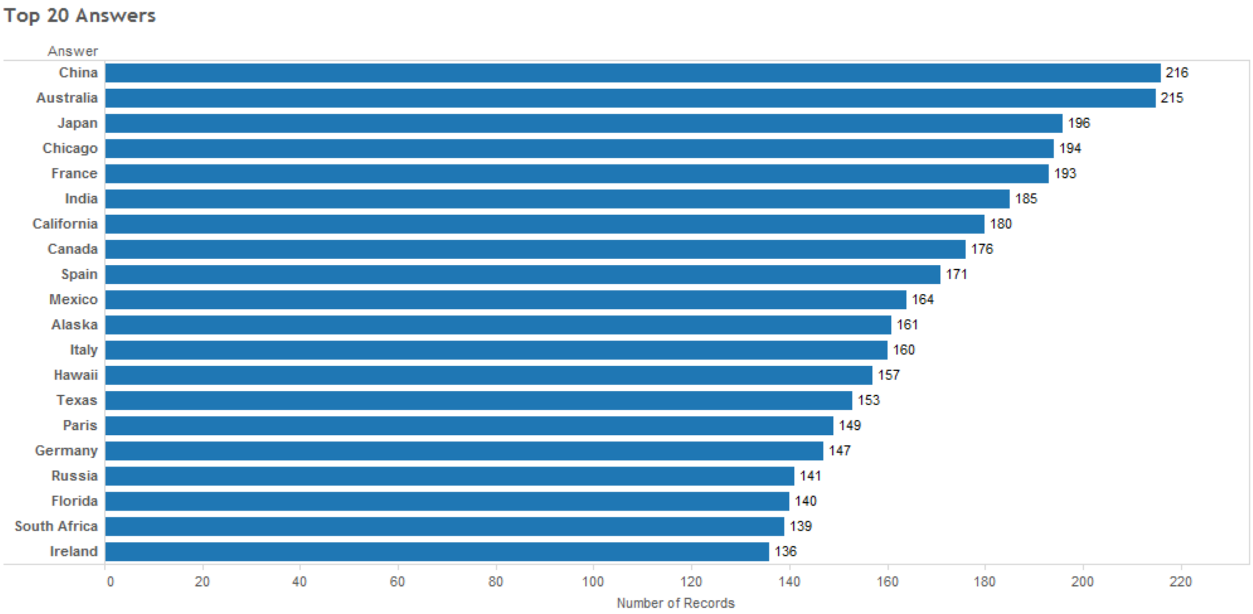
Overall top 20 categories by #records by airdate



The trend of sum of Number of Records for Air Date (group) broken down by Category. The marks are labeled by sum of Number of Records. The data is filtered on Round, which keeps Double Jeopardy!, Final Jeopardy!, Jeopardy! and Tiebreaker. The view is filtered on Category, which keeps 20 of 27,995 members.

Step 4: The questions are harder to analyze since they obviously need a lot of contextual information and after all they form the crux of the show. One option was to try and categorize these questions into higher-level categories using NLP but that path took us nowhere. We soon discovered that reducing 28K categories would take a lot more time that we have.

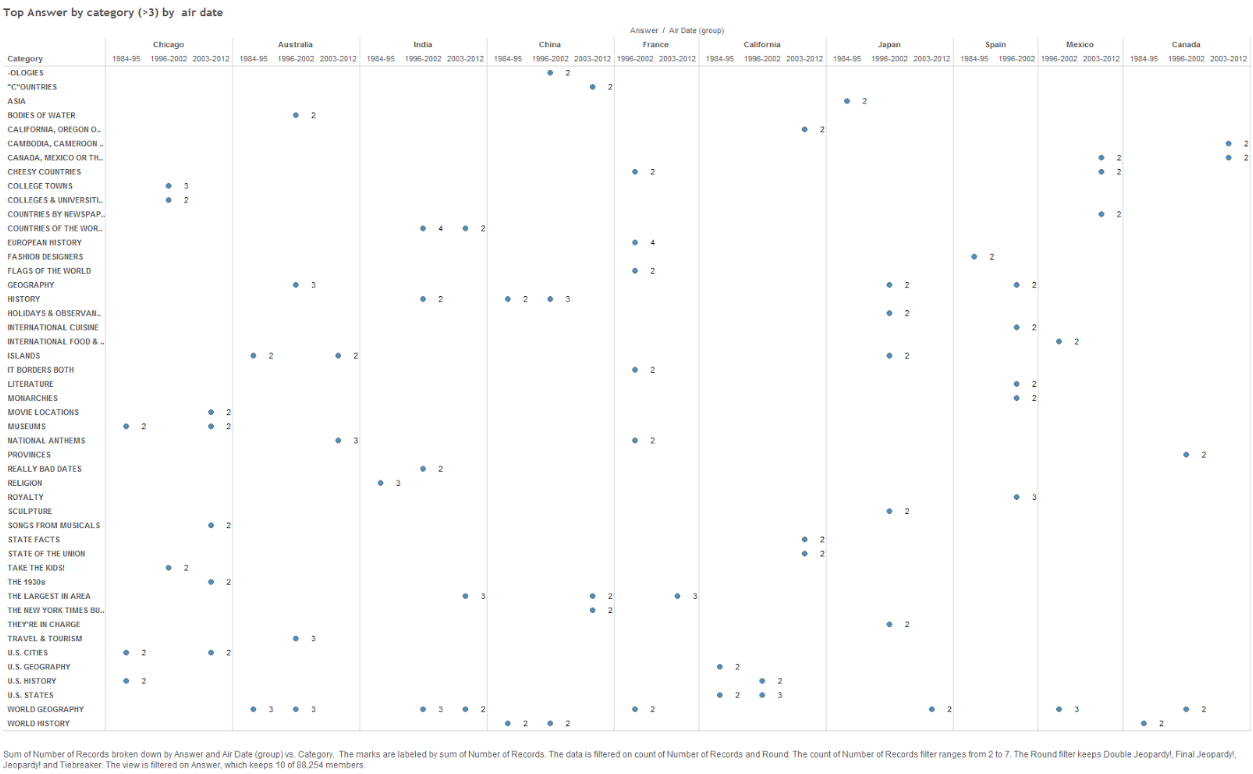
The next meaningful step was then to look at the answers. We didn't really know what to expect but the answers definitely were interesting.



Sum of Number of Records for each Answer. The marks are labeled by sum of Number of Records. The data is filtered on Round, which keeps Double Jeopardy!, Final Jeopardy!, Jeopardy! and Tiebreaker. The view is filtered on Answer, which keeps 20 of 88,254 members.

The standout commonality - everything again seems to be related to Geography. Knowing your countries may not be a bad place to start.

Step 5: Now that we have a set of answers that seem to be popular, we wanted to check how these questions and categories map to each other. Even though, these form a very minor fraction of the total number of distinct answers (88K of them!) This graph just shows those categories that had the answer feature at least 2 times and as we can see, there are not so many. If we filter that in include categories where each question has at least 4 occurrences, this further drops to just 2 category-answer combinations. There are a large number of specific categories where each answer comes up once, not all obviously related to geography. Therefore even if all the answers seem heavily pointing towards Geograph, we cannot pick that out unless we group the categories.

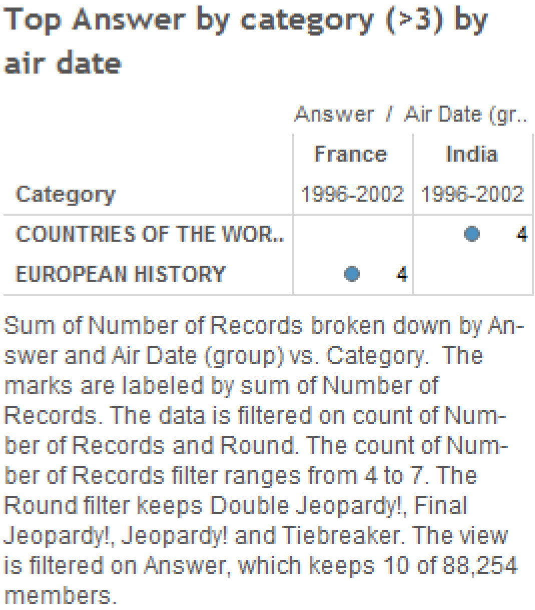


Top Answer by category (>3) by air date

Answer / Air Date (group)

	Chicago			Australia			India			China			France		California			Japan			Spain		Mexico		Canada		
Category	1984-95	1996-2002	2003-2012	1984-95	1996-2002	2003-2012	1984-95	1996-2002	2003-2012	1984-95	1996-2002	2003-2012	1996-2002	2003-2012	1984-95	1996-2002	2003-2012	1984-95	1996-2002	2003-2012	1984-95	1996-2002	1996-2002	2003-2012	1984-95	1996-2002	2003-2012
-OLOGIES											● 2																
"C"OUNTRIES												● 2															
ASIA																		● 2									
BODIES OF WATER					● 2																						
CALIFORNIA, OREGON O..																	● 2										
CAMBODIA, CAMEROON ..																											● 2
CANADA, MEXICO OR TH..																								● 2			● 2
CHEESY COUNTRIES													● 2											● 2			
COLLEGE TOWNS		● 3																									
COLLEGES & UNIVERSITI..		● 2																									
COUNTRIES BY NEWSPAP..																								● 2			
COUNTRIES OF THE WOR..								● 4	● 2																		
EUROPEAN HISTORY													● 4														
FASHION DESIGNERS																					● 2						
FLAGS OF THE WORLD													● 2														
GEOGRAPHY					● 3														● 2			● 2					
HISTORY								● 2		● 2	● 3																
HOLIDAYS & OBSERVAN..																			● 2								
INTERNATIONAL CUISINE																						● 2					
INTERNATIONAL FOOD & ..																							● 2				
ISLANDS				● 2		● 2													● 2								
IT BORDERS BOTH													● 2														
LITERATURE																						● 2					
MONARCHIES																						● 2					
MOVIE LOCATIONS			● 2																								
MUSEUMS	● 2		● 2																								
NATIONAL ANTHEMS						● 3							● 2														
PROVINCES																										● 2	
REALLY BAD DATES								● 2																			
RELIGION							● 3																				
ROYALTY																						● 3					
SCULPTURE																			● 2								
SONGS FROM MUSICALS			● 2																								
STATE FACTS																	● 2										
STATE OF THE UNION																	● 2										
TAKE THE KIDS!		● 2																									
THE 1930s			● 2																								
THE LARGEST IN AREA									● 3			● 2		● 3													
THE NEW YORK TIMES BU..												● 2															
THEY'RE IN CHARGE																			● 2								
TRAVEL & TOURISM					● 3																						
U.S. CITIES	● 2		● 2																								
U.S. GEOGRAPHY															● 2												
U.S. HISTORY	● 2															● 2											
U.S. STATES															● 2	● 3											
WORLD GEOGRAPHY				● 3	● 3			● 3	● 2				● 2							● 2			● 3			● 2	
WORLD HISTORY										● 2	● 2														● 2		

Sum of Number of Records broken down by Answer and Air Date (group) vs. Category. The marks are labeled by sum of Number of Records. The data is filtered on count of Number of Records and Round. The count of Number of Records filter ranges from 2 to 7. The Round filter keeps Double Jeopardy!, Final Jeopardy!, Jeopardy! and Tiebreaker. The view is filtered on Answer, which keeps 10 of 88,254 members.



Top Answer by category (>3) by air date

Answer / Air Date (gr..

	France	India
Category	1996-2002	1996-2002
COUNTRIES OF THE WOR..		● 4
EUROPEAN HISTORY	● 4	

Sum of Number of Records broken down by Answer and Air Date (group) vs. Category. The marks are labeled by sum of Number of Records. The data is filtered on count of Number of Records and Round. The count of Number of Records filter ranges from 4 to 7. The Round filter keeps Double Jeopardy!, Final Jeopardy!, Jeopardy! and Tiebreaker. The view is filtered on Answer, which keeps 10 of 88,254 members.

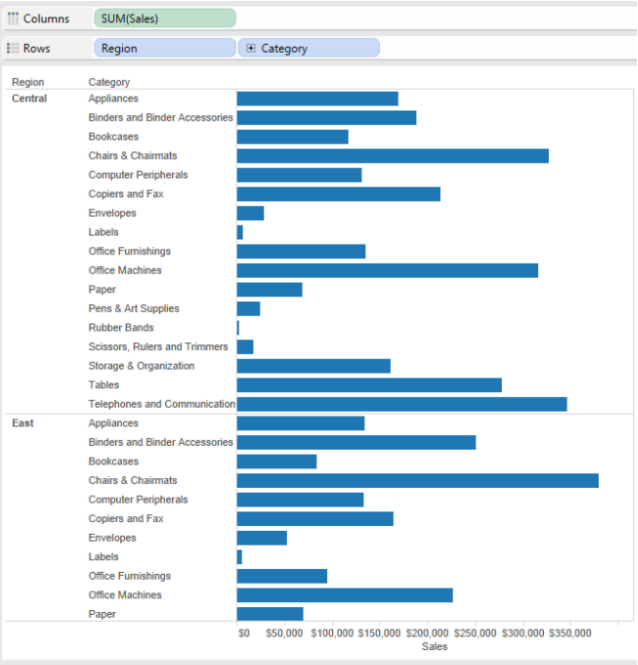
Conclusion: At this point, we concluded that while Geography seems to be a strong contender, unless we can group categories together we may not be able to identify an overarching favorite.

Another realization was that maybe we should have subdivided the dataset before analyzing these questions. The most sensible division being that based on the rounds in the game. So our next hypothesis was: **Is there a difference between Single Jeopardy, Double Jeopardy and Final Jeopardy in terms of the content?**

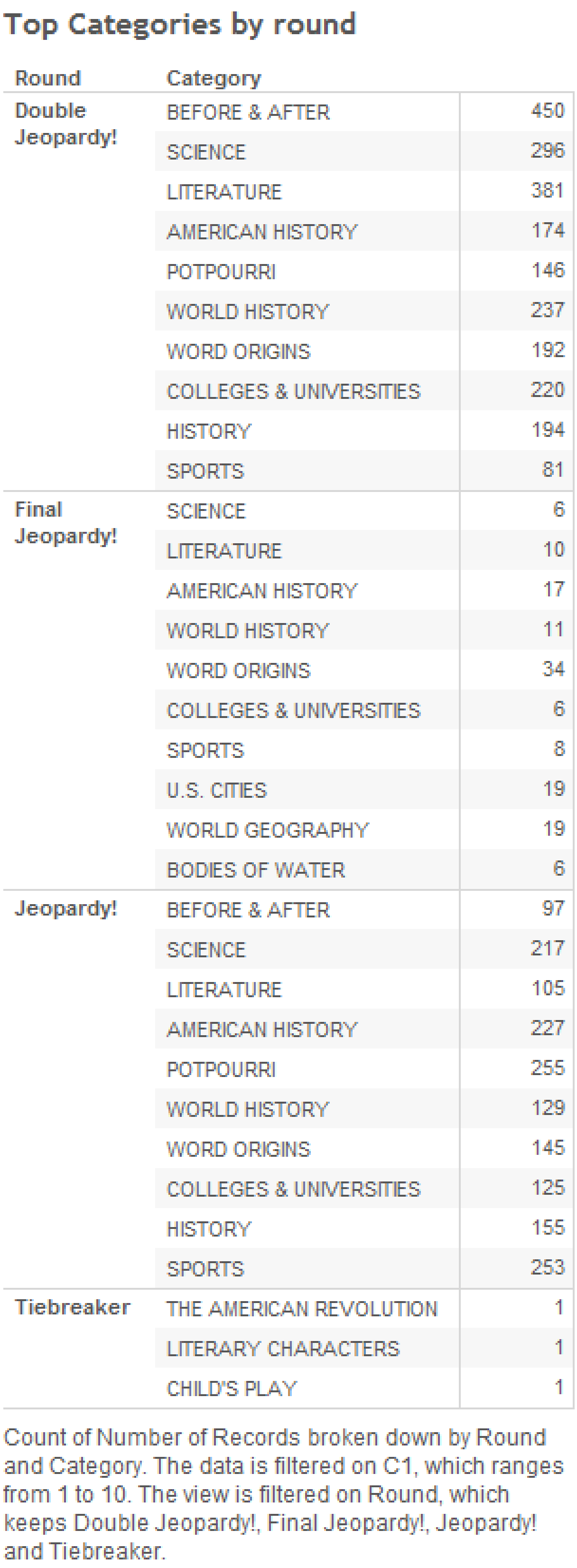
Step 1: Identifying what were the many categories for each round of the show.

This turned out to be a lot harder because of the technical challenges involved. Initially to filter by "Round", we tried to use the simple filter option available for each category. A query of the nature - "Filter by Top 10 on Count of Number of Records in Round" to try and select the top ten question and then control this using a user filter to select the round. However, this does not work. It just retains the same questions as was used to define the query and on user modification, it filters out of the same list.

A little research identified the right method to perform an operation of this sort. It was to create a **"Calculated field"** and use that to create a ranking of elements within a subcategory. A very useful **related work** to understand this was this article written in the Tableau Knowledge base http://kb.tableau.com/articles/knowledgebase/finding-top-n-within-category [4] which talks about "Finding the Top N Within a Category" where they use a parallel of identifying Sales within a region. The corollary in this case being identifying the top categories within a round.



In our case, we were able to identify the top categories by each Round in Jeopardy.



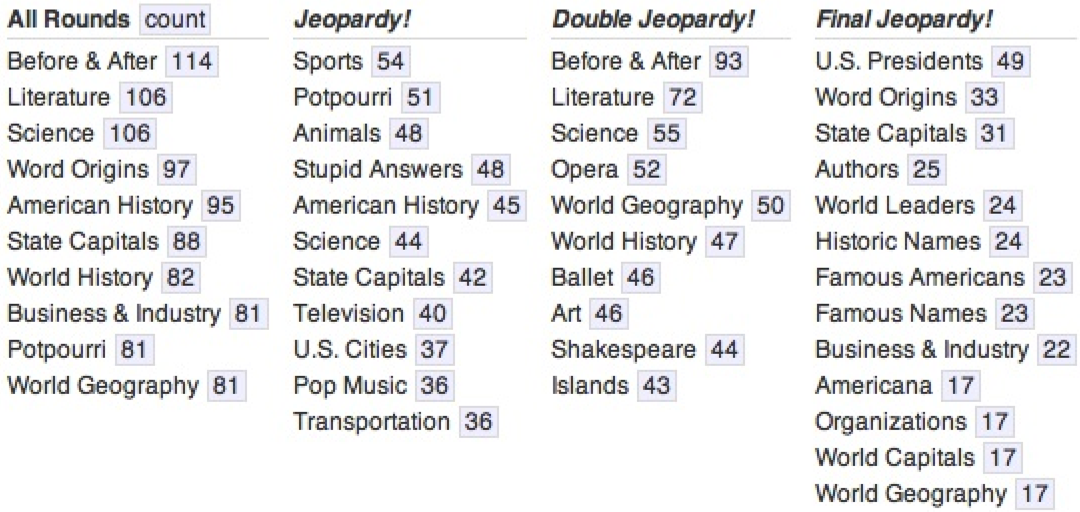
Top Categories by round

Round	Category	
Double Jeopardy!	BEFORE & AFTER	450
	SCIENCE	296
	LITERATURE	381
	AMERICAN HISTORY	174
	POTPOURRI	146
	WORLD HISTORY	237
	WORD ORIGINS	192
	COLLEGES & UNIVERSITIES	220
	HISTORY	194
	SPORTS	81
Final Jeopardy!	SCIENCE	6
	LITERATURE	10
	AMERICAN HISTORY	17
	WORLD HISTORY	11
	WORD ORIGINS	34
	COLLEGES & UNIVERSITIES	6
	SPORTS	8
	U.S. CITIES	19
	WORLD GEOGRAPHY	19
	BODIES OF WATER	6
Jeopardy!	BEFORE & AFTER	97
	SCIENCE	217
	LITERATURE	105
	AMERICAN HISTORY	227
	POTPOURRI	255
	WORLD HISTORY	129
	WORD ORIGINS	145
	COLLEGES & UNIVERSITIES	125
	HISTORY	155
	SPORTS	253
Tiebreaker	THE AMERICAN REVOLUTION	1
	LITERARY CHARACTERS	1
	CHILD'S PLAY	1

Count of Number of Records broken down by Round and Category. The data is filtered on C1, which ranges from 1 to 10. The view is filtered on Round, which keeps Double Jeopardy!, Final Jeopardy!, Jeopardy! and Tiebreaker.

Step 2: To add to the categories, we found the top answers in each category.

Insights from the answers and questions by category were heavily biased by an article we came across on *Slate*, which had an article done on Jeopardy http://www.slate.com/articles/arts/culturebox/2011/02/ill_take_jeopardy_trivia_for_200_alex.html

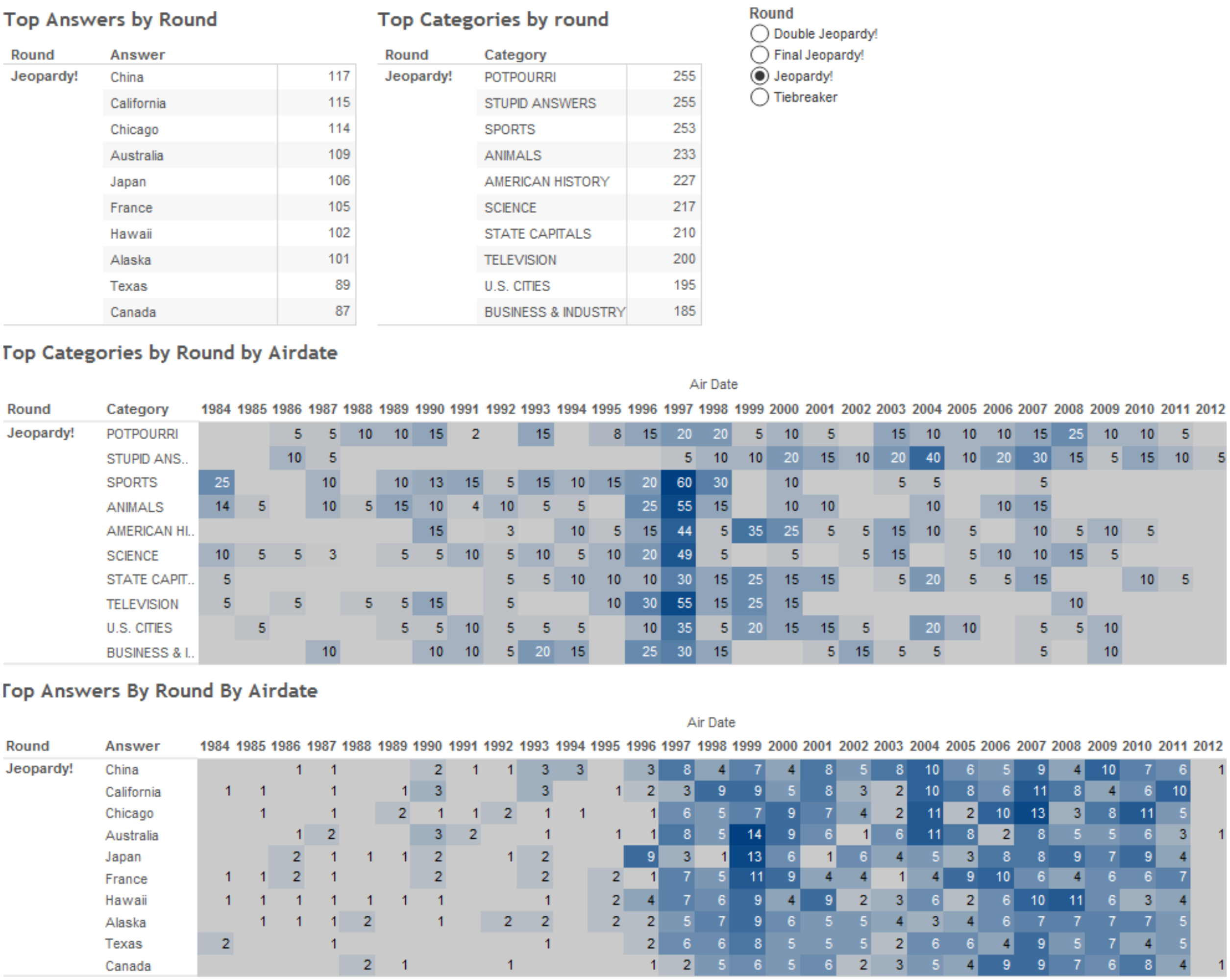


All Rounds	count	***Jeopardy!***		***Double Jeopardy!***		***Final Jeopardy!***	
Before & After	114	Sports	54	Before & After	93	U.S. Presidents	49
Literature	106	Potpourri	51	Literature	72	Word Origins	33
Science	106	Animals	48	Science	55	State Capitals	31
Word Origins	97	Stupid Answers	48	Opera	52	Authors	25
American History	95	American History	45	World Geography	50	World Leaders	24
State Capitals	88	Science	44	World History	47	Historic Names	24
World History	82	State Capitals	42	Ballet	46	Famous Americans	23
Business & Industry	81	Television	40	Art	46	Famous Names	23
Potpourri	81	U.S. Cities	37	Shakespeare	44	Business & Industry	22
World Geography	81	Pop Music	36	Islands	43	Americana	17
		Transportation	36			Organizations	17
						World Capitals	17
						World Geography	17

The analysis showed how a certain theme could be identified in each round. To quote the article, "Using this method of analysis, a portrait of the first round starts to emerge—and it looks like grade school. Double Jeopardy!, meanwhile, is more like college, with a touch of the yacht club while Final Jeopardy! screams patriotism, with a dash of diplomacy and a dearth of science"

Step 3: We also wanted to analyse how categories varied with time. We felt that our earlier analysis showed a clustering of category repetition during a specific time frame and we wanted to check if this extended into each round as well. Screenshots of this can be seen on the dashboard below.

Step 4: Then a dashboard showing the variation with airdate. This is based on a very similar dashboard done before which I came across on Tableau which had used heat maps to analyze time based data. Given the similarity, we decide to talk about this process in our report instead of using it in our final visualization. We've hosted it on the public Tableau Server though : https://public.tableau.com/views/Jeopardy_0/Dashboard1?:embed=y&:showTabs=y&:display_count=yes



Top Answers by Round

Round	Answer	
Jeopardy!	China	117
	California	115
	Chicago	114
	Australia	109
	Japan	106
	France	105
	Hawaii	102
	Alaska	101
	Texas	89
	Canada	87

Top Categories by round

Round	Category	
Jeopardy!	POTPOURRI	255
	STUPID ANSWERS	255
	SPORTS	253
	ANIMALS	233
	AMERICAN HISTORY	227
	SCIENCE	217
	STATE CAPITALS	210
	TELEVISION	200
	U.S. CITIES	195
	BUSINESS & INDUSTRY	185

Round
○ Double Jeopardy!
○ Final Jeopardy!
● Jeopardy!
○ Tiebreaker

Top Categories by Round by Airdate

Air Date

Round	Category	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Jeopardy!	POTPOURRI			5	5	10	10	15	2		15		8	15	20	20	5	10	5		15	10	10	10	15	25	10	10	5	
	STUPID ANS..			10	5										5	10	10	20	15	10	20	40	10	20	30	15	5	15	10	5
	SPORTS	25			10		10	13	15	5	15	10	15	20	60	30		10			5	5			5					
	ANIMALS	14	5		10	5	15	10	4	10	5	5		25	55	15		10	10			10		10	15					
	AMERICAN HI..							15		3		10	5	15	44	5	35	25	5	5	15	10	5		10	5	10	5		
	SCIENCE	10	5	5	3		5	5	10	5	10	5	10	20	49	5		5		5	15		5	10	10	15	5			
	STATE CAPIT..	5								5	5	10	10	10	30	15	25	15	15		5	20	5	5	15			10	5	
	TELEVISION	5		5		5	5	15		5			10	30	55	15	25	15								10				
	U.S. CITIES		5				5	5	10	5	5	5		10	35	5	20	15	15	5		20	10		5	5	10			
	BUSINESS & I..				10			10	10	5	20	15		25	30	15			5	15	5	5			5		10			

Top Answers By Round By Airdate

Air Date

Round	Answer	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Jeopardy!	China			1	1			2	1	1	3	3		3	8	4	7	4	8	5	8	10	6	5	9	4	10	7	6	1
	California	1	1		1		1	3			3		1	2	3	9	9	5	8	3	2	10	8	6	11	8	4	6	10	
	Chicago		1		1		2	1	1	2	1	1		1	6	5	7	9	7	4	2	11	2	10	13	3	8	11	5	
	Australia			1	2			3	2		1		1	1	8	5	14	9	6	1	6	11	8	2	8	5	5	6	3	1
	Japan			2	1	1	1	2		1	2			9	3	1	13	6	1	6	4	5	3	8	8	9	7	9	4	
	France	1	1	2	1			2			2		2	1	7	5	11	9	4	4	1	4	9	10	6	4	6	6	7	
	Hawaii	1	1	1	1	1	1	1			1		2	4	7	6	9	4	9	2	3	6	2	6	10	11	6	3	4	
	Alaska		1	1	1	2		1		2	2		2	2	5	7	9	6	5	5	4	3	4	6	7	7	7	7	5	
	Texas	2			1						1			2	6	6	8	5	5	5	2	6	6	4	9	5	7	4	5	
	Canada					2	1			1				1	2	5	6	5	6	2	3	5	4	9	9	7	6	8	4	1

Exploring the wagers data

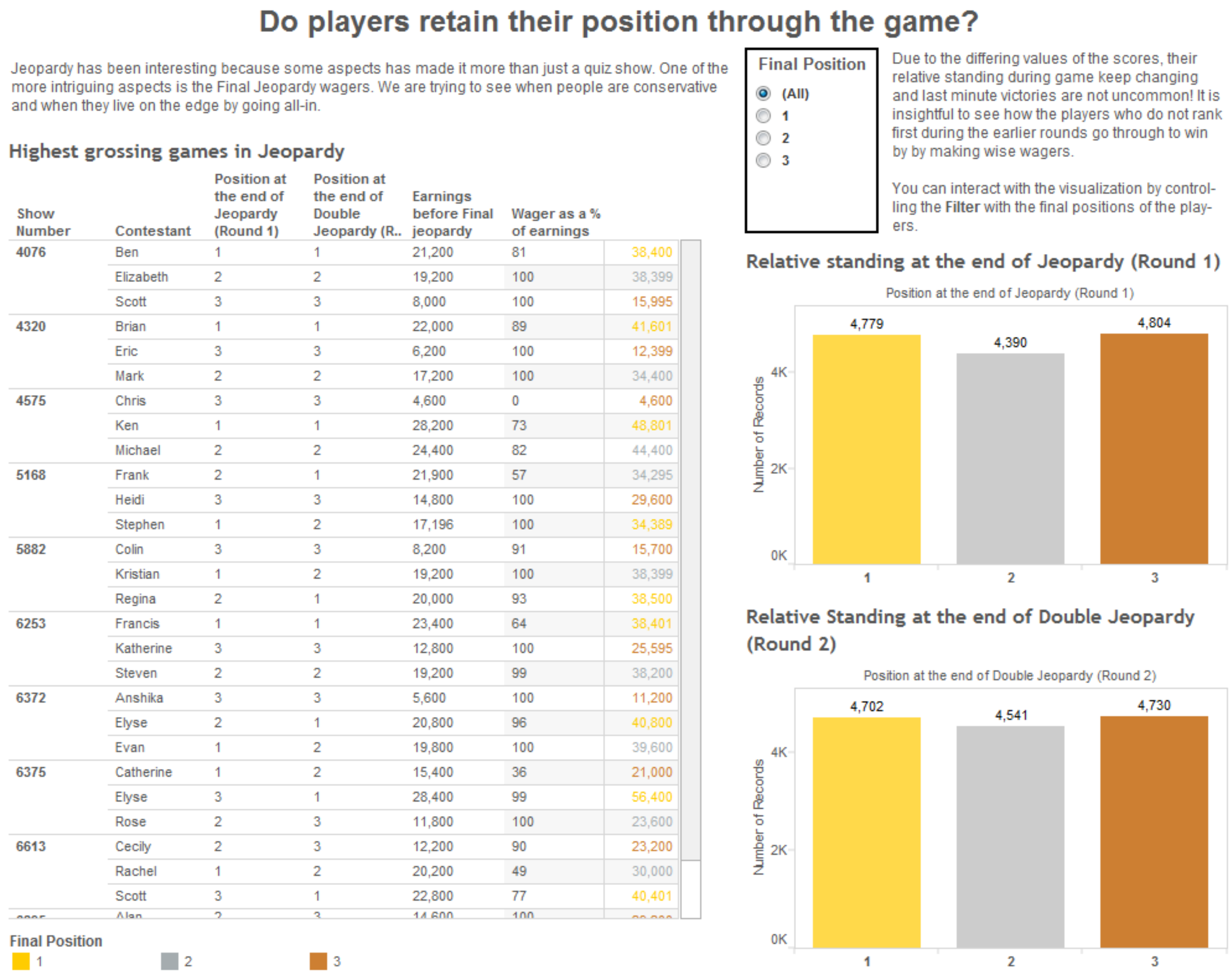
This was a two stage process. We started with some manipulation in Python to aggregate the data in the format we needed. The original data had the progress by each question, it had a cumulative score per contestant after each question in the game. After transforming it get the total by round and the individual wagers we explored the data in Tableau.

Step1: The relative standing of the contestants in each Single Jeopardy and Double

This was mainly with the intention of identifying whether there is a clear and dominant winner through the game. We felt that the game was structured that there were enough points to cause fluctuations in the player standings. We found that the data supported this assumption. We used this to also develop a D3 visualization and show this flow. After initially trying this on Tableau, we found it harder to customize to our needs and switched to D3.

Step 2: We wanted to see how player's standing changed at each game, the wagers they made based on their positions and their final standing after it. We looked at different lists at this point, the top twenty games where the players had made highest individual earnings, the games where there was a minimum difference between the top two places and finally we settled on games where there was a maximum total earnings. These games showed more fluctuation in the relative standing of the players during the game since everyone did well.

Step 3: Creating a meaningful dashboard that showed this data in an easy to read manner with sufficient explanation to allow users to understand the content and interact with it.



Do players retain their position through the game?

Jeopardy has been interesting because some aspects has made it more than just a quiz show. One of the more intriguing aspects is the Final Jeopardy wagers. We are trying to see when people are conservative and when they live on the edge by going all-in.

Highest grossing games in Jeopardy

Show Number	Contestant	Position at the end of Jeopardy (Round 1)	Position at the end of Double Jeopardy (R..	Earnings before Final jeopardy	Wager as a % of earnings	
4076	Ben	1	1	21,200	81	38,400
	Elizabeth	2	2	19,200	100	38,399
	Scott	3	3	8,000	100	15,995
4320	Brian	1	1	22,000	89	41,601
	Eric	3	3	6,200	100	12,399
	Mark	2	2	17,200	100	34,400
4575	Chris	3	3	4,600	0	4,600
	Ken	1	1	28,200	73	48,801
	Michael	2	2	24,400	82	44,400
5168	Frank	2	1	21,900	57	34,295
	Heidi	3	3	14,800	100	29,600
	Stephen	1	2	17,196	100	34,389
5882	Colin	3	3	8,200	91	15,700
	Kristian	1	2	19,200	100	38,399
	Regina	2	1	20,000	93	38,500
6253	Francis	1	1	23,400	64	38,401
	Katherine	3	3	12,800	100	25,595
	Steven	2	2	19,200	99	38,200
6372	Anshika	3	3	5,600	100	11,200
	Elyse	2	1	20,800	96	40,800
	Evan	1	2	19,800	100	39,600
6375	Catherine	1	2	15,400	36	21,000
	Elyse	3	1	28,400	99	56,400
	Rose	2	3	11,800	100	23,600
6613	Cecily	2	3	12,200	90	23,200
	Rachel	1	2	20,200	49	30,000
	Scott	3	1	22,800	77	40,401
[illegible]	Alan	2	3	14,600	100	[illegible]

Final Position

🟨 1 ⬜ 2 🟧 3

Final Position

● (All)
○ 1
○ 2
○ 3

Due to the differing values of the scores, their relative standing during game keep changing and last minute victories are not uncommon! It is insightful to see how the players who do not rank first during the earlier rounds go through to win by by making wise wagers.

You can interact with the visualization by controlling the **Filter** with the final positions of the players.

Relative standing at the end of Jeopardy (Round 1)

Relative Standing at the end of Double Jeopardy (Round 2)

This dashboard underwent a lot of changes after feedback from a lot of people. Initially, the columns headers and text were not descriptive enough and people spent too much time trying to understand the message conveyed.The color combinations were not ideal. The bar charts were ugly. The filter was difficult to spot. Overall, it had a lot of problems.

We got some more feedback from Prof. Marti after I had actually tried to fix a lot of the previous issues. All this feedback was immensely helpful. This was one of the most challenging parts of our entire project - identify a good and interesting narrative from a lot of factual details. Unlike many project, this didn't have a natural story to it so a lot of our exploration was a process of trying to identify content what people would like to engage with. We came to a conclusion that if it was something very specific to the show, then it does not have much value. By exploring this story, we are trying to find aspects of the game that are more generic. Contestant behavior. Jeopardy is now just an example to demonstrate this aspect of human behavior, where contestants wager differently based on their position. This is still presented in a way that we talk about Jeopardy because that's what the overall narrative is about but it has some value to readers outside of the game.

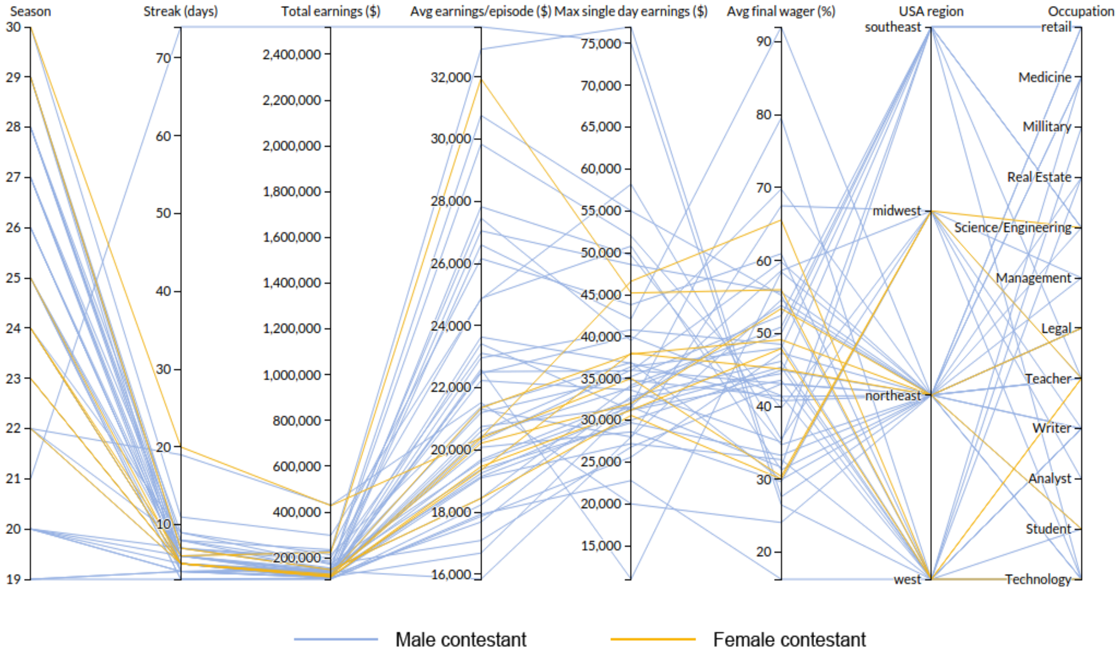
A lot of the practices that went into designing this board came from another article we read on the Tableau knowledge base
http://kb.tableau.com/articles/knowledgebase/best-practices-designing-vizes-and-dashboards
[5]
This article talks about good practices when designing a dashboards and a lot of these principles can be tied back to the design principles we studied in class.

The color choices, which was something we wanted to keep consistent throughout our narrative, was based on the visualization wheel described by *Cairo* in the book "the functional art". We wanted a balance between originality and familiarity and given how this was linked to the filter which is not immediately obvious, we wanted to use a color combination that people would immediately recognize and there is nothing more well known than the gold-silver -bronze combination. From our feedback, we used this source http://www.brandigirlblog.com/2012/11/why-do-some-color-schemes-work-and-others-dont.html [6] to identify which colors went with each other to get the right shades for all the other components so that they match these colors.

Parallel Coordinate Visualization

We wanted to explore the top 50 contestants who had the most earnings in the game's history. We also wanted to map these contestants to some of the features like winning streak length, maximum earnings in a day, their gender, the region where they came from and their occupation.

We built a parallel coordinate visualization where each contestant was represented by a line and the different axes were the attributes like total earnings, streak length in days, average earnings per episode, Maximum earnings in a single day, USA Region, Occupation and gender.

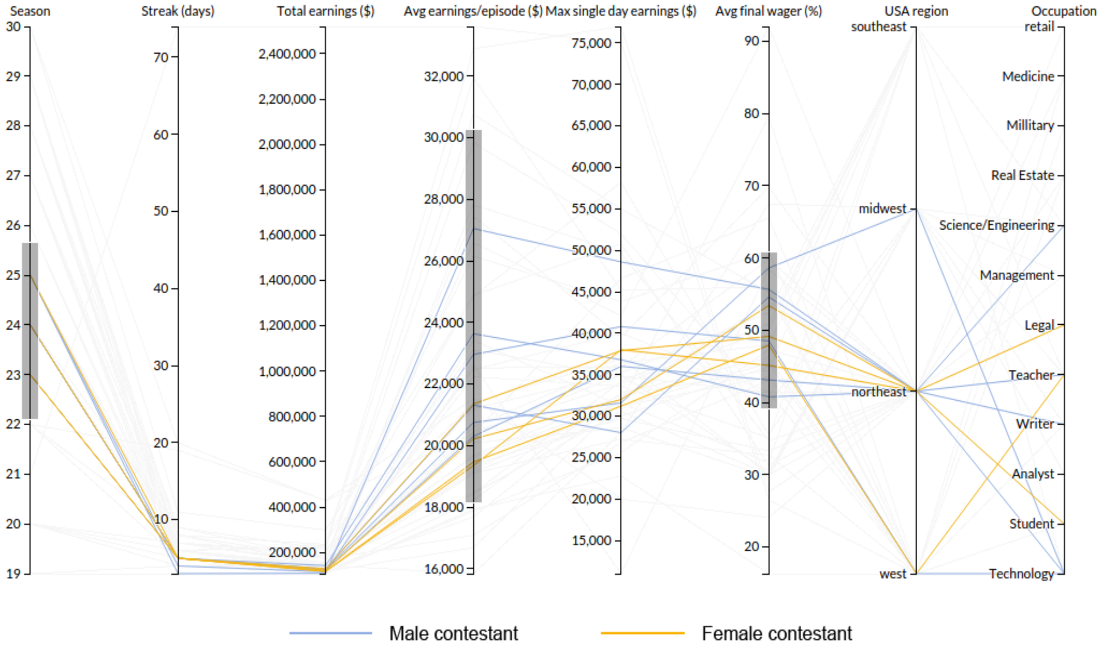


The main skeleton of the parallel coordinates visualization has been inspired from a similar work here "http://bl.ocks.org/mbostock/1341021" [7]. This visualized the various aspects of a car like economy, power, weight etc which we extended to our case in Jeopardy! using the various characteristics of the contestants as axes.

Brushing and Linking

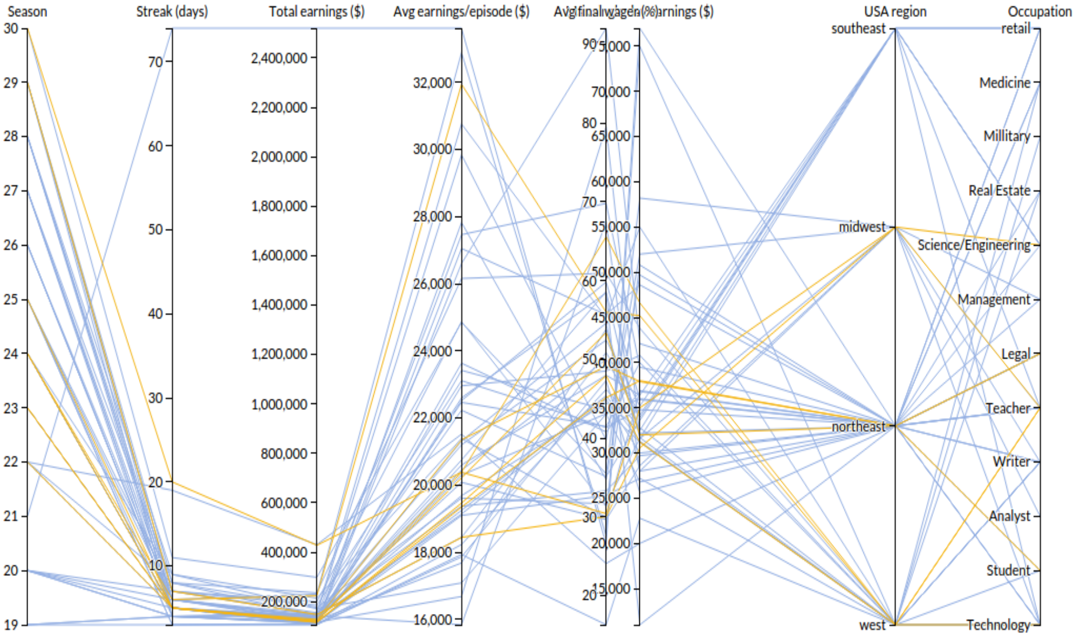
We incorporated brushing and linking into the parallel coordinate visualization. Brushing is a very effective technique for specifying an explicit focus during information visualization. The user actively marks subsets of the data-set as being especially interesting, for example, by using a brush-like interface element. If used in conjunction with multiple linked views, brushing can enable users to understand correlations across multiple dimensions. This allows the user to filter the data based on particular values for multiple features at the same time. For example in the picture below, we have used brushing to narrow down on the season(23 -25), Average earnings per episode(18000$ to 30000$) and Average final wager percentage(40 - 60%) all at the same time. This gives the user a lot of flexibility to play and narrow down on the data.

We referred to this paper "Angular Brushing of Extended Parallel Coordinates" by Helwig Hauser, Florian Ledermann, and Helmut Doleisch [8] which discusses the extensions of the parallel coordinates visualizations specifically Brushing and axes re-ordering. The article discusses how Brushing and axes ordering allow the users to explore the data better and identify connections or patterns easily. We tried to incorporate these features into our visualization so that we could provide more flexibility and functionality to the users.



Axes reordering

We added a feature in the parallel coordinate visualization that allows the user to re -order the axes around by just dragging them into a particular gap (between two other axes) and the axes would automatically position themselves.



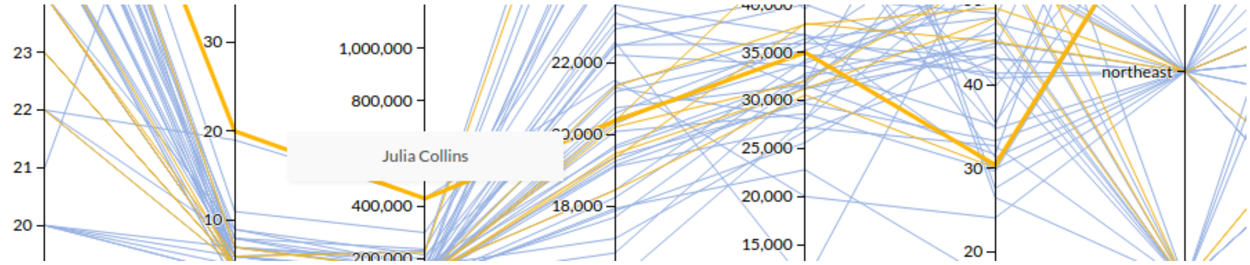
Ordinal/Categorical Variables

In our parallel coordinate visualization we have used categorical variables for two axes - USA region and Occupation. Initially we had coded the categorical variables to number and decided to convey what categories number corresponded to through text, however based on the user feedback that we got, that there was a lot of interpretation required to map the numbers to categories especially for the axes that had a larger number of categories(Occupation).
One of the bug in the parallel coordinates that was introduced after we used ordinal variables in place of numeric variables was that brushing does not work for the axes that display ordinal variables. We have not been able to resolve this as of now but we are working on this and try to look into the reference mentioned below.

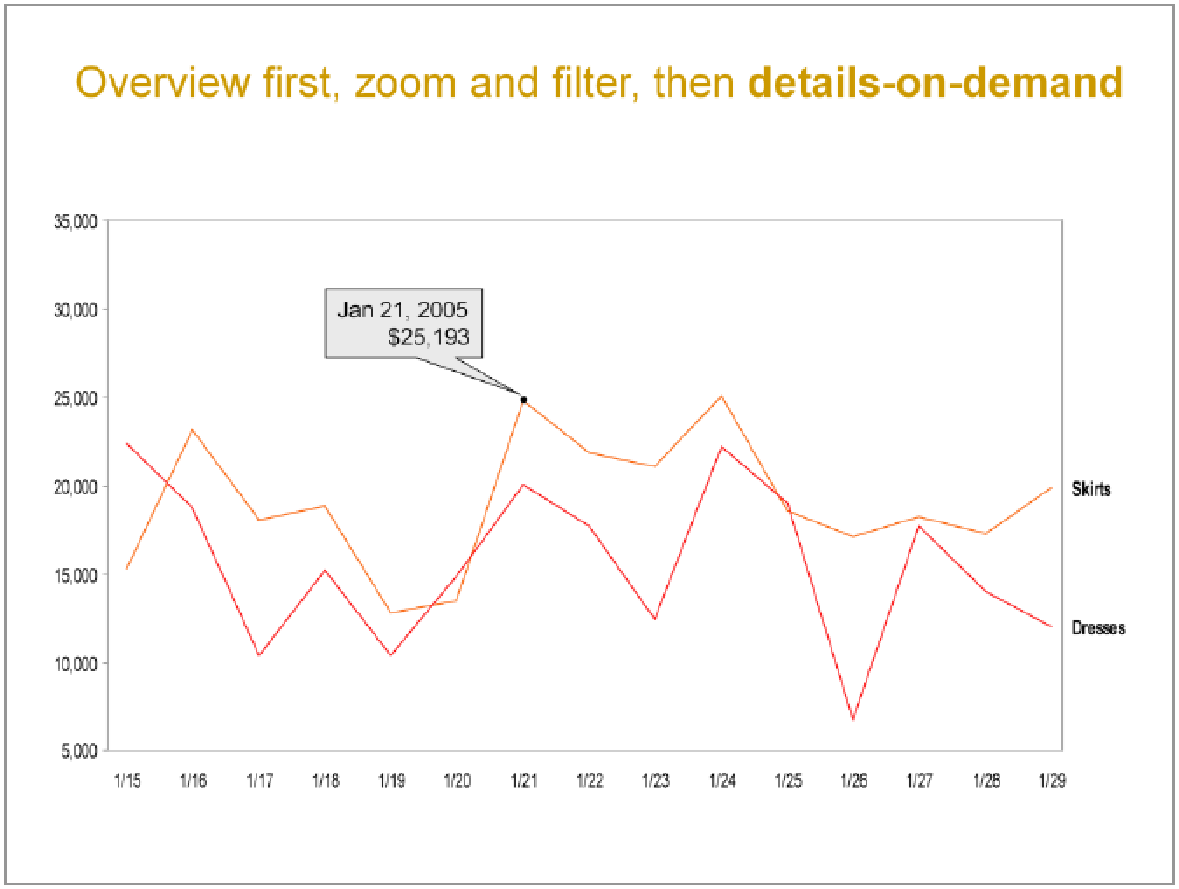
"http://bl.ocks.org/mbostock/4349509" [9], an example where there is focus to include brushing in ordinal and categorical axes. We would like to look into this example in future and try to apply it to our visualization which currently doesn't allow brushing on the axes with ordinal variables.

Tooltip

We added a tooltip identify each contestant in the line graph. Since there were a lot of interesting outliers or extremes in the data, we decided you use a tooltip so that the user can directly relate to the player when they see something interesting. Another important tweak that we added was to highlight the entire path of the contestant that was selected, so that the user could focus on that particular contestant across all axes. Here in the figure below, we can see that the user wanted to know more about the top yellow line(yellow -> female) ie the female contestant who has the highest earnings. Not only does the tooltip show up displaying the name of the contestant but the user can focus on the values of that contestant across all axes because of the line getting highlighted.



Stephen Few talks about Details on Demand on pg 116 of his book "Now you see it" [10]. Few talks about "the need of precise details from time to time that cannot be seen just by looking at the data. He proposes a way to access that detail without departing from the rich visual environment called Detail on Demand, a feature that allows the details to become visible when needed and disappear otherwise thereby avoiding the cluttering of the screen and distraction". We took a cue from this and tried to implement this for most of the visualizations like the line charts, parallel coordinates and the calendar heatmap heat map.



Gender Divide

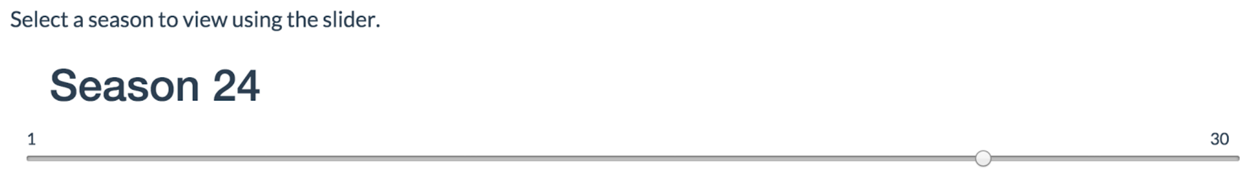
During the mid project presentation, we got some feedback on including demographic data for the contestants and that might reveal some of the interesting facts about the data. Once we included gender, occupation and USA region, we observed that there was a strong skew in favor of males - there were 42 males and only 8 females in the top 50 contestants. We decided to highlight this issue under the gender divide section and removed gender as an axis and used colors (used across all visualizations to represent genders) to distinguish the gender of a contestant.

Ken Jennings

Ken Jennings is undisputedly the best player to have appeared on Jeopardy! Once we plotted the top 50 contestants we saw that Ken Jennings skewed the two axes - total winnings at 2,500,000$ compared to 450,000$ next and 74 win streaks as compared to 20 next. Including Ken Jennings actually compressed the rest of the data to the lower end of the axes. We plan to use a logarithmic scale for these axes in the future so that we can have better arrangement of data for these axes.

Slider

Working with 30 seasons of Jeopardy! data challenged us to think of effective ways of displaying the data. In many applications, such as with a scatter plot, showing too many data points can detract from the goal of a visualization. Filtering, as Stephen Few defines, "is the act of reducing the data that we're viewing to a subset of what's currently there" (p. 64).



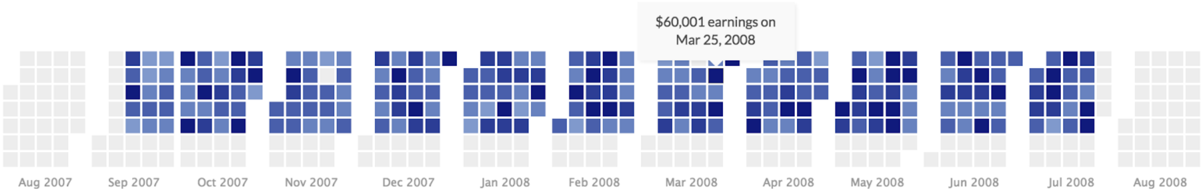
Select a season to view using the slider.

A slider, because of its horizontal layout, which can be indicative of a time-based relationship, was a logical choice for us to filter the Jeopardy! data. This is created using an HTML range object. In addition to including the minimum and maximum values on the left and right sides of the slider, respectively, we added a season label above it. This works using two functions. First, the user gets immediate feedback on the position of the slider with the use of the oninput event. This updates the season label based on the slider position, letting users know which season they'd select were they to release the slider. The second function makes use of the onchange event, which actually triggers a change in the data filter. For more information on these function, see the onchange vs. oninput for Range Sliders article.

Heat Map

Drawing inspiration from the exploratory data analysis in Tableau, specifically the heat map on the categories and answers, we chose to use a heat map to provide an overview of the earnings data. The heat map was created using an API, though we slightly modified the JavaScript for our purposes.

We first started by exploring the data by year, which is a natural way to consider time series data. We then realized that the earnings information covered two seasons and that it would show a gap during the summer when the show is off the air. Based on this, we decided to visualize data at the season level. We also made the decision to separate each calendar month. This was intended to make it easier for users to target a specific date. The labels are also meant to aid in that process.

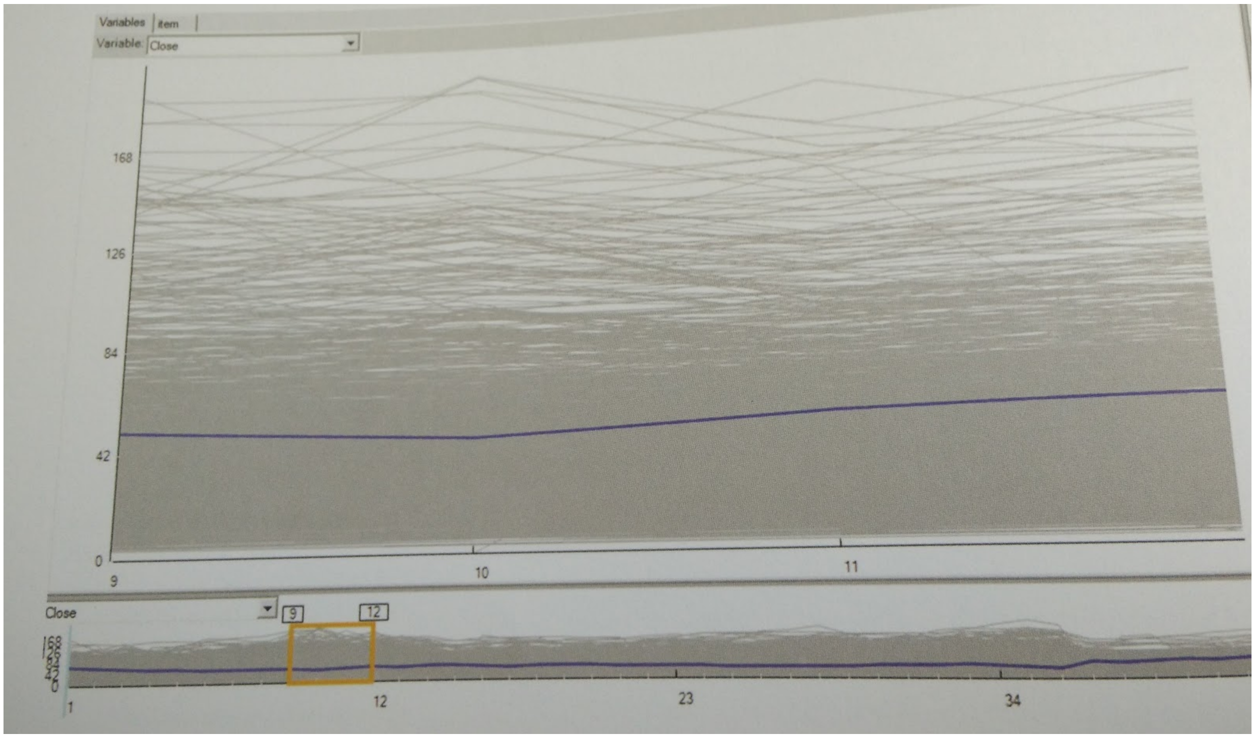


There are seven squares in (almost) every column of the heat map, each representing a day of the week. The gray squares are ones without data. This can be for either the summer, the weekends, or days with missing data. The two rows of gray squares at the bottom help frame the blue-shaded ones.

Because the heat map covers an entire season, we decided to structure it by month. In the heat map, each square corresponds to a particular date and the darkness of the shading indicated the earnings level. As Stephen Few describes, "there are times when we need to see precise details that can't be discerned in the visualization" (p. 87)[12]. This is the case with the heat map. It is intended to show a general overview of the earnings in a particular season. To Few's point, we added a tooltip, or "pop-up box," as Few calls it, to the heat map that shows additional information when a square is moused over. The squares show the relative earnings on a particular episode and are based on the following range: `[10000, 25000, 40000, 55000]`. For example, there is a shade for less than 10,000, between 10,000 and 25,000, etc.

Linking Line Chart

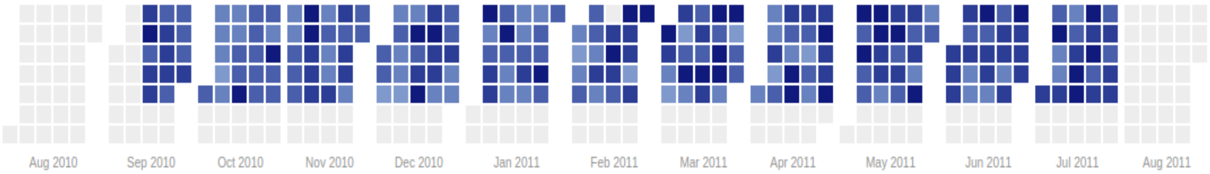
Stephen Few talks about "Focus and context Together" on pg 113 in his book "Now you see it" [13], where he says that it is very easy to lose context of the bigger picture when you are focussing on the details. The solution he proposes is generically called "Focus + Context" . He says " When we are focussing on details, the whole doesn't need to be visible in high resolution, but we need to see where the details that we are focussing on lie in the bigger picture and how they relate to it". Few proposes the use of zooming as a way to achieve this by focussing on a particular portion while the picture of the whole is available to us.

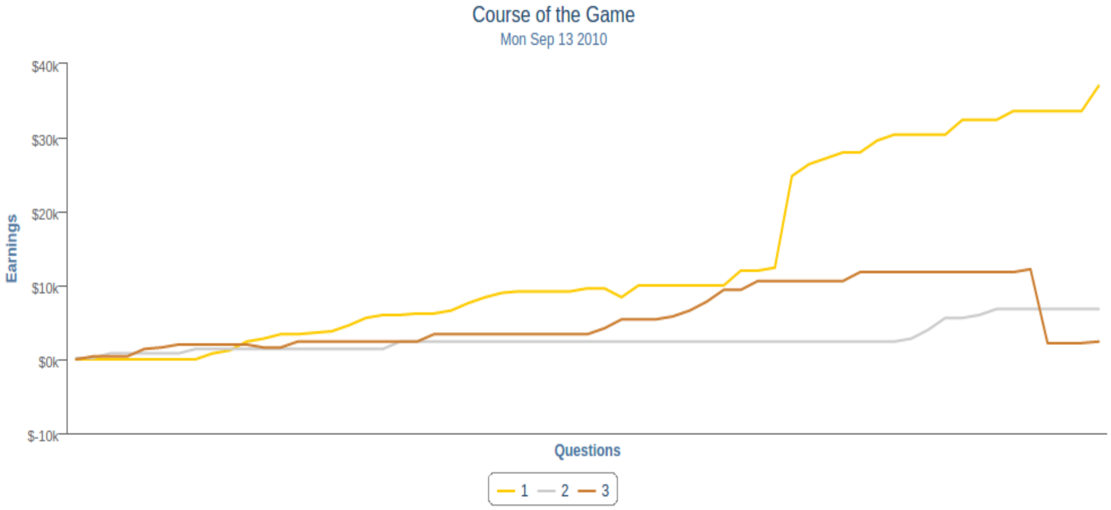


We implemented this concept of Few's using the linking line charts. After we built a calendar heat map to display the total/average earnings for each of the episodes for a particular season and integrating it with a slider that allows the user to switch through the different seasons. We decided to link a line graph to the episode heat map so that when you click on a particular episode on the heatmap, a line chart populates that shows the entire flow of the game on that particular episode. The line chart was developed using highcharts and has three lines that are populated for each of the three contestants. The color for the three lines in the line chart carry over from the ones we have used to denote positions across the entire site(gold for 1st, silver for 2nd and bronze for 3rd). There are effectively 61 points that join to form the line. Each of these points represent the earnings of the contestants after the question jeopardy 1 to 30 , double jeopardy 1 to 30 and the final jeopardy.

We integrated a tool tip for the line charts based on Few's "Detail on Demand" feature that he mentions on pg 116 in his book "Now you see it",   to show scores after a particular question when you over at a point on one of the lines in the highcharts. We plan to add the contestant names in the future so that the user can get a better understanding of who was playing. We will have to merge multiple datasets on a primary key to include the contestant name into this.

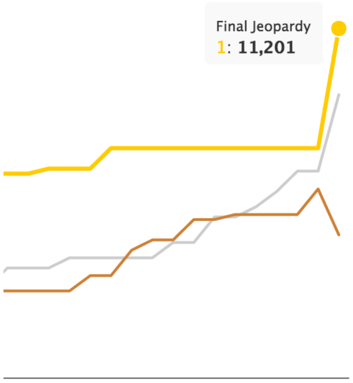
The toughest challenge that we faced in integrating the line chart was to filter the data based on date corresponding to the square that was clicked on the season heat map. We also defaulted to the first episode of the season when the slider was moved to change to a different season.



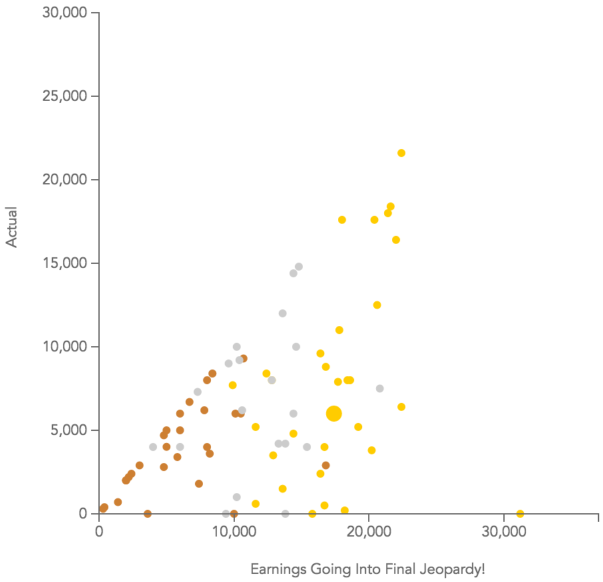


Scatter Plot

One facet of the game we were interested in exploring was the Final Jeopardy! wagers. As shown below, the biggest potential for earnings and positions to change occur in the Final Jeopardy! round. To explore how contestants make wagers, we created a scatter plot of wagers against earnings.

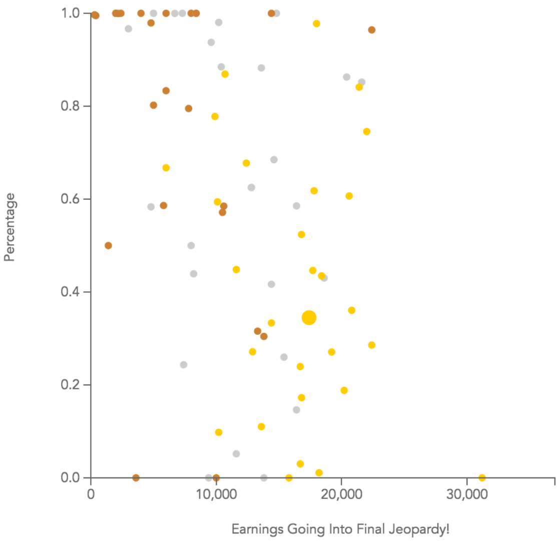


On the x-axis, we plot the earnings going into the Final Jeopardy! round. The circles are colored based on the contestants' position at this point in time. We decided that this was more indicative of the type of wagers contestants would make.



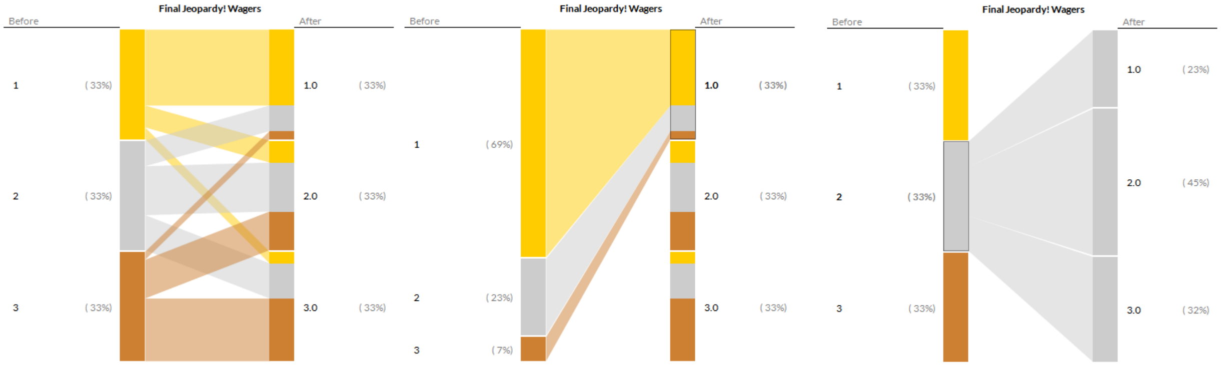
For example, we can see that, in general, contestants in third place, identified by the bronze color, wager all or close to all of their earnings. An additional feature we added was the ability to discriminate between objects by increasing the size of the circle that is moused over. As Cairo explains, "the brain groups similar objects (the rectangles of same size and tone) and separates them from those that look different" (p. 114).

While the first place contestants placed higher wagers, they place the lowest wagers on a percentage basis. This is why we included a dropdown menu for users to select between an "actual" and "percentage" view. (The circle that was called out in the previous screenshot is called out again below.)



We also thought significantly about transitions with this data. When transitioning between the actual and percentage views, we decided to keep the circles the same colors, but change their size, making them slightly smaller during the transition and then returning to their regular size. This was done to signal that the circles in each view correspond to the same data point (contestants). In contrast, when changing between seasons, the circles change to black on the transition so that it's clear that the circles refer to other data points.

Bipartite Graph



The idea and concept for this visualization was directly related to the exploratory analysis we had done earlier. This visualization was effective in showing how the players changed their positions before and after final jeopardy. While not the most common occurrence, nevertheless sufficient to talk about.

This visualization was strongly based on this D3 example http://bl.ocks.org/NPashaP/9796212 [15]

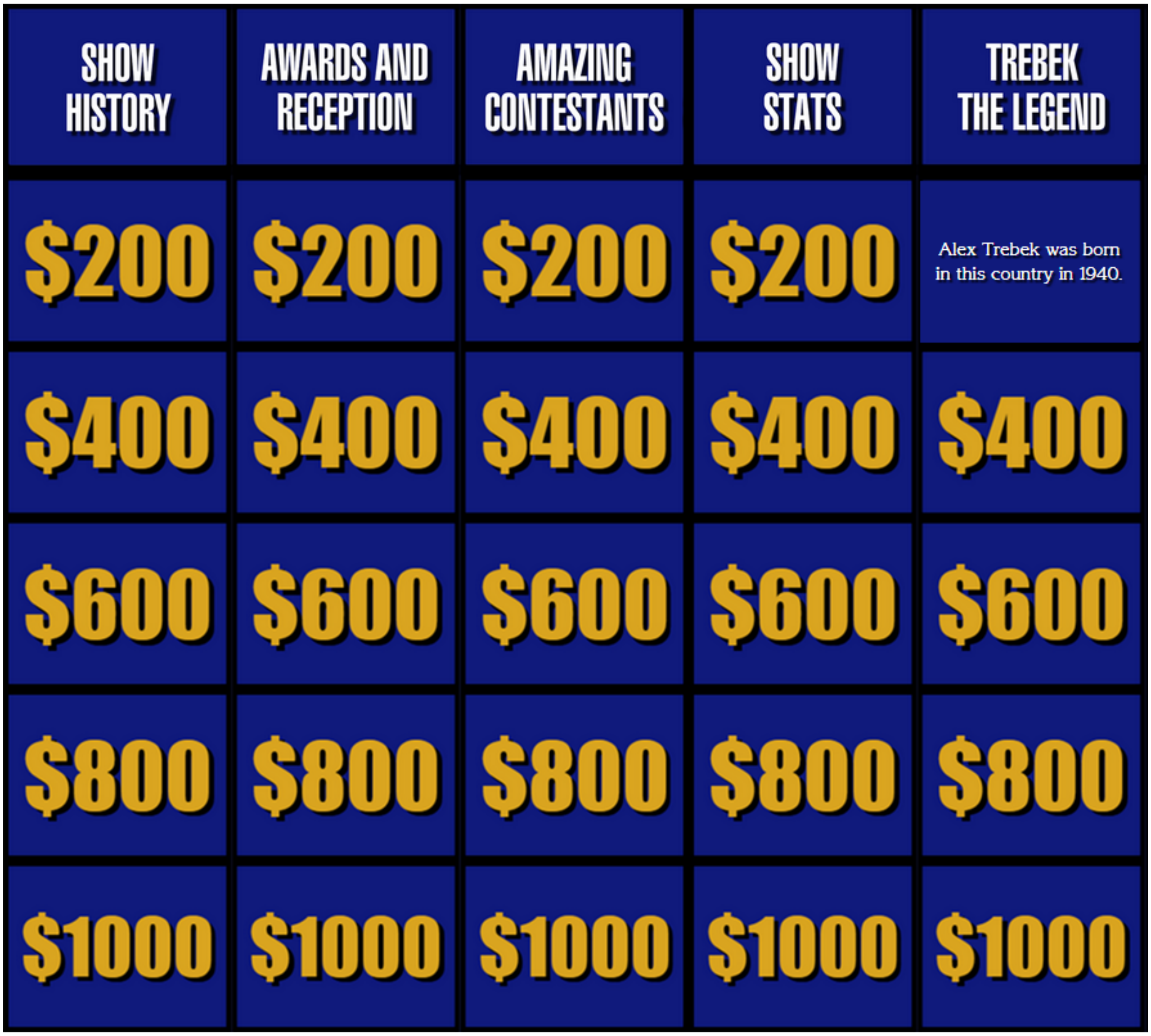
We used this as reference to create the visualization using data cleaned and generated and Python and fed directly into this backend. This shows the percentage in each category before and after the main action - the final wager.. It re-enforces the ideas discussed in the exploration dashboard which comes earlier in the narrative.

Jeopardy Game Board

After people reached the end of our visualization, we wanted to give them something fun to play with at the end of the game. It is also an interactive activity that we thought would be entertaining for groups of people while we presented. Most of the facts used on the board were collected from Wikipedia, blogs and other articles:

http://mashable.com/2014/03/30/jeopardy-facts/
http://www.grandparents.com/food-and-leisure/did-you-know/jeopardy-trivia
http://en.wikipedia.org/wiki/Jeopardy!

We wrote the questions ourselves, devised a way to categorize them and used Justinmind, an interface prototyping tool, to make the game interactive. We looked up articles on the fonts used in Jeopardy (http://fontsinuse.com/uses/5507/jeopardy-game-show), found them, downloaded them and used them on this board (as well as the title of the webpage) to make it all look authentic.



SHOW HISTORY	AWARDS AND RECEPTION	AMAZING CONTESTANTS	SHOW STATS	TREBEK THE LEGEND
$200	$200	$200	$200	Alex Trebek was born in this country in 1940.
$400	$400	$400	$400	$400
$600	$600	$600	$600	$600
$800	$800	$800	$800	$800
$1000	$1000	$1000	$1000	$1000

Links

http://people.ischool.berkeley.edu/~anand_rajagopal/Jeopardy/

Bibliography (Only Related work):

1. Cairo's (p.51) visualization wheel
2. Few visualization project on government spending (Pg 305)
3. Project by Otto and Marie Meurath about Home and Factory Weaving in England (Cairo p.72)
4. Tableau knowledge base on Calculated Fields http://kb.tableau.com/articles/knowledgebase/finding-top-n-within-category

5. Tableau knowledge base on Dashboard Design Best Practices http://kb.tableau.com/articles/knowledgebase/best-practices-designing-vizes-and-dashboards
6. Choosing the right colors http://www.brandigirlblog.com/2012/11/why-do-some-color-schemes-work-and-others-dont.html
7. Parallel Coordinates D3 http://bl.ocks.org/mbostock/1341021
8. Hochheiser, Harry, and Ben Shneiderman. "Dynamic query tools for time series data sets: timebox widgets for interactive exploration." *Information Visualization* 3.1 (2004): 1-18.
9. Brushing for ordinal variables http://bl.ocks.org/mbostock/4349509
10. Data on Demand, pg 116, "Now you see it" by Stephen Few
11. Calendar Heat Map http://kamisama.github.io/cal-heatmap/v2/
12. Stephen Few - Level of precision in detail [Page 86]
13. Focus and context together, pg 113, "Now you see it" by Stephen Few
14. D3 Scatterplot http://bl.ocks.org/WilliamQLiu/bd12f73d0b79d70bfbae
15. Bipartite graph http://bl.ocks.org/NPashaP/9796212

Work Distribution

A majority of all our work was done together working in groups trying to accomplish a task, much like the peer programming and development practices we developed in class. Given that, it was very hard to try and find a way to split this data. Please find a table below that gives an approximate distribution - but is not completely indicative of every member's work in ensuring that the product satisfied the goals we had for our work. We've all contributed more in terms of non quantifiable work of testing, usability and design changes.

	Juan	**Anand**	**Anubhav**	**Joshua**
Concept and Ideas	25%	25%	25%	25%
Exploratory Data Analysis using Tableau	0%	100%	0%	0%
Gender Analysis	0%	0%	0%	100%
Text Descriptions	25%	25%	25%	25%
Web Scraping	80%	0%	20%	0%
Data Cleaning and Manipulation	50%	20%	20%	10%
User Testing	25%	25%	25%	25%

Gender and Contestant Infographics	0%	0%	0%	100%
Contestant Appearance and Wins by Gender Highcharts Line Graph	0%	0%	10%	90%
Champions parallel coordinates	0%	0%	90%	10%
Player position tableau	0%	100%	0%	0%
Seasons slider + episode heatmap	90%	0%	10%	0%
Episode Dynamics Line Graph	20%	0%	80%	0%
Final Wager Risk Scatter Plot	80%	0%	20%	0%
Final Wager Bipartite	100%	0%	0%	0%
Jeopardy Game Board	0%	0%	0%	100%
Website Integration	10%	60%	10%	20%
Report Writing	25%	25%	25%	25%

Thank you.