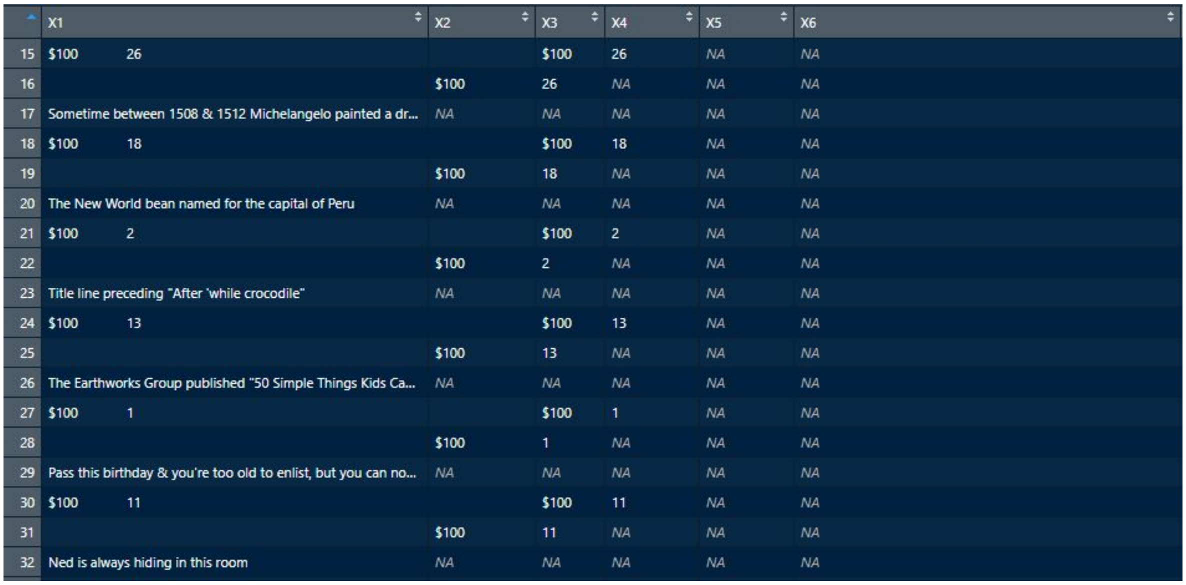
Problem Set 5 - ECON 5253

Thomas Mondry

March 2, 2022

1 Web scraping – J! Archive

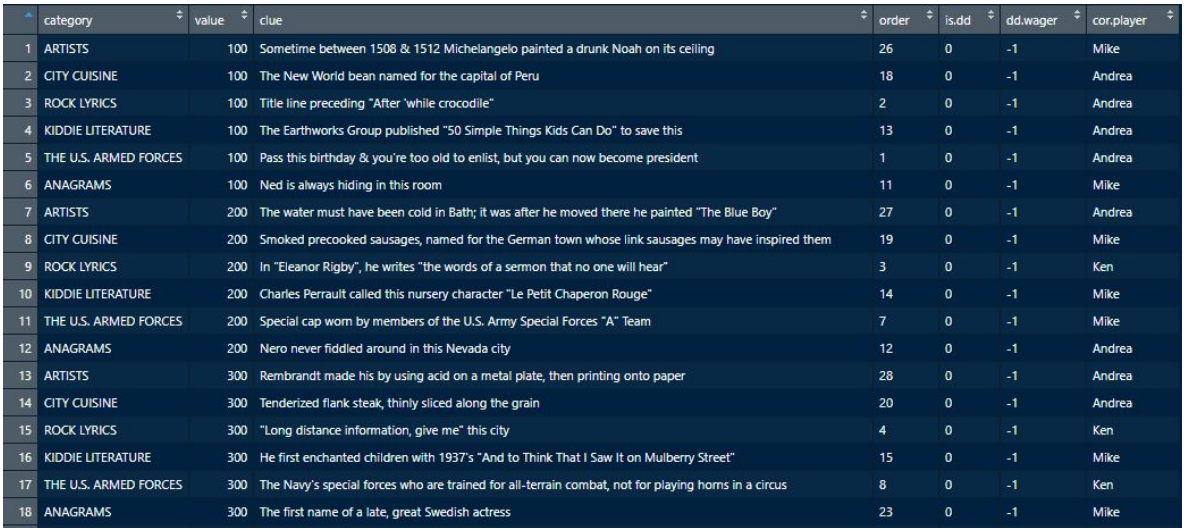
For this problem, I wanted to try out manual web scraping from the suboptimally-formatted J! Archive website to see whether this would be a feasible approach to data acquisition for my final project. The site is flashy and clearly tailored toward the experience of someone viewing the games one at a time in browser, with correct responses & some other information usually only displayed on mouseover. While the way `rvest` parses the tables is hard to read and somewhat unpredictable (see Figure 1), it is consistent (from what I can tell) among different rounds from different games, meaning it is possible to programmatically transform the data from many games into a readable format. I gave this a shot here, deciding that the best approach would be to start entirely from scratch with an empty data frame of relevant information, which I then populated one row at a time by looking in the right places in the raw data. One trick I found was that while the absolute location of the information of interest for each clue within the entire table doesn't follow a consistent pattern, its position relative to a respective `NA` cell is consistent, so the location of the relevant information can easily be obtained after these `NA` cells are located.



	X1	X2	X3	X4	X5	X6
15	$100 26		$100	26	NA	NA
16		$100	26	NA	NA	NA
17	Sometime between 1508 & 1512 Michelangelo painted a dr...	NA	NA	NA	NA	NA
18	$100 18		$100	18	NA	NA
19		$100	18	NA	NA	NA
20	The New World bean named for the capital of Peru	NA	NA	NA	NA	NA
21	$100 2		$100	2	NA	NA
22		$100	2	NA	NA	NA
23	Title line preceding "After 'while crocodile"	NA	NA	NA	NA	NA
24	$100 13		$100	13	NA	NA
25		$100	13	NA	NA	NA
26	The Earthworks Group published "50 Simple Things Kids Ca...	NA	NA	NA	NA	NA
27	$100 1		$100	1	NA	NA
28		$100	1	NA	NA	NA
29	Pass this birthday & you're too old to enlist, but you can no...	NA	NA	NA	NA	NA
30	$100 11		$100	11	NA	NA
31		$100	11	NA	NA	NA
32	Ned is always hiding in this room	NA	NA	NA	NA	NA

Figure 1: Part of the raw table of clues, as parsed by `rvest`

The end results here were a table of clue-level information for one round of one game (Figure 2); and, more importantly, a reproducible & sufficiently efficient process for scraping the data from two different URLs, parsing the correct elements into an R data frame, and transforming the raw data into a usable format. I expect that this process can easily be wrapped into a function which takes the ID of the game on the archive – fortunately, the URLs on the site appear to be straightforward and consistent – and returns complete tables of each clue from each game. Since the data contained on the site encompasses most of the quantifiable information about a game (with the exception of buzzer times and other information which is very difficult to obtain), I believe my process will be sufficient for



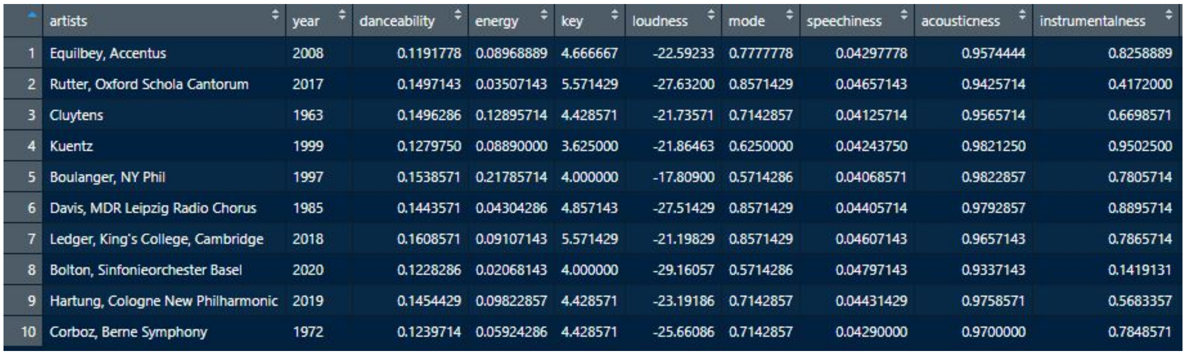
	category	value	clue	order	is.dd	dd.wager	cor.player
1	ARTISTS	100	Sometime between 1508 & 1512 Michelangelo painted a drunk Noah on its ceiling	26	0	-1	Mike
2	CITY CUISINE	100	The New World bean named for the capital of Peru	18	0	-1	Andrea
3	ROCK LYRICS	100	Title line preceding "After 'while crocodile"	2	0	-1	Andrea
4	KIDDIE LITERATURE	100	The Earthworks Group published "50 Simple Things Kids Can Do" to save this	13	0	-1	Andrea
5	THE U.S. ARMED FORCES	100	Pass this birthday & you're too old to enlist, but you can now become president	1	0	-1	Andrea
6	ANAGRAMS	100	Ned is always hiding in this room	11	0	-1	Mike
7	ARTISTS	200	The water must have been cold in Bath; it was after he moved there he painted "The Blue Boy"	27	0	-1	Andrea
8	CITY CUISINE	200	Smoked precooked sausages, named for the German town whose link sausages may have inspired them	19	0	-1	Mike
9	ROCK LYRICS	200	In "Eleanor Rigby", he writes "the words of a sermon that no one will hear"	3	0	-1	Ken
10	KIDDIE LITERATURE	200	Charles Perrault called this nursery character "Le Petit Chaperon Rouge"	14	0	-1	Mike
11	THE U.S. ARMED FORCES	200	Special cap worn by members of the U.S. Army Special Forces "A" Team	7	0	-1	Mike
12	ANAGRAMS	200	Nero never fiddled around in this Nevada city	12	0	-1	Andrea
13	ARTISTS	300	Rembrandt made his by using acid on a metal plate, then printing onto paper	28	0	-1	Andrea
14	CITY CUISINE	300	Tenderized flank steak, thinly sliced along the grain	20	0	-1	Andrea
15	ROCK LYRICS	300	"Long distance information, give me" this city	4	0	-1	Ken
16	KIDDIE LITERATURE	300	He first enchanted children with 1937's "And to Think That I Saw It on Mulberry Street"	15	0	-1	Mike
17	THE U.S. ARMED FORCES	300	The Navy's special forces who are trained for all-terrain combat, not for playing horns in a circus	8	0	-1	Ken
18	ANAGRAMS	300	The first name of a late, great Swedish actress	23	0	-1	Mike

Figure 2: Part of the cleaned data table for the Jeopardy round which aired 05/23/1991

tackling any data-driven question about the game which I may want to investigate for my final project, although there are a few more pieces of information which I'd like to integrate into the process.

2 API – Spotify data

For this problem, I pulled data from the Spotify API using the `spotifyr` package. I decided to take advantage of the audio features which the Spotify API provides in order to quantitatively compare several recordings of the same piece of music – Fauré's *Requiem*. I used the API to find 10 popular albums that consist solely of the *Requiem*, then obtained audio features for each track of each album, and finally averaged these audio features for each album in order to compare the recordings quantitatively.



	artists	year	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness
1	Equilbey, Accentus	2008	0.1191778	0.08968889	4.666667	-22.59233	0.7777778	0.04297778	0.9574444	0.8258889
2	Rutter, Oxford Schola Cantorum	2017	0.1497143	0.03507143	5.571429	-27.63200	0.8571429	0.04657143	0.9425714	0.4172000
3	Cluytens	1963	0.1496286	0.12895714	4.428571	-21.73571	0.7142857	0.04125714	0.9565714	0.6698571
4	Kuentz	1999	0.1279750	0.08890000	3.625000	-21.86463	0.6250000	0.04243750	0.9821250	0.9502500
5	Boulanger, NY Phil	1997	0.1538571	0.21785714	4.000000	-17.80900	0.5714286	0.04068571	0.9822857	0.7805714
6	Davis, MDR Leipzig Radio Chorus	1985	0.1443571	0.04304286	4.857143	-27.51429	0.8571429	0.04405714	0.9792857	0.8895714
7	Ledger, King's College, Cambridge	2018	0.1608571	0.09107143	5.571429	-21.19829	0.8571429	0.04607143	0.9657143	0.7865714
8	Bolton, Sinfonieorchester Basel	2020	0.1228286	0.02068143	4.000000	-29.16057	0.5714286	0.04797143	0.9337143	0.1419131
9	Hartung, Cologne New Philharmonic	2019	0.1454429	0.09822857	4.428571	-23.19186	0.7142857	0.04431429	0.9758571	0.5683357
10	Corboz, Berne Symphony	1972	0.1239714	0.05924286	4.428571	-25.66086	0.7142857	0.04290000	0.9700000	0.7848571

Figure 3: Table with a few features for some recordings of the Fauré *Requiem*

Glancing at the results, there doesn't appear to be a lot of major variation in the features. This is to be expected – everyone is fundamentally performing the same music. A couple of notes:

- It appears at first that the Nadia Boulanger/NY Phil recording is a big outlier, as its loudness and energy scores are much higher than the other recordings; however, a quick listen reveals that this is likely because it is a live recording, and a very old one at that (the performance is from 1962), so recording technology didn't allow for the removal of audience coughs and other ambient noises.

- For some reason, the Bolton/Sinfonieorchester Basel recording has a surprisingly low instrumentalness score of 0.142, where most of the others are in the 0.6-0.9 range. It sounds like this recording was made on period string instruments, which have a slightly "harsher" sound –

maybe the algorithm that generates the audio features interpreted this sound as electronic, but it doesn't seem like the difference should be so pronounced.