

Title: Learnings from the spoonacular recipe database of over 365000 recipes

Name: Ben Tankus

Date: 8 February 2021

Problem statement: What is the best recipe, and what factors contribute to recipe quality?

Americans have a hard time cooking at home. If a high quality (spoon score) recipe can be predicted and suggested based on a few relevant factors I believe we can help Americans cook at home more. This will save money, improve health, and many other potential positive outcomes.

The questions I will answer here are as follows:

1. Is it worth getting more pans dirty? (Is there a positive correlation between spoonscore and equipment count)
2. What factor contributes the most to spoon score?
3. Are vegan recipes more expensive?

Obtain

The same data set was used for the group project. The following text in the *Obtain* section is from the initial report:

To determine what the project topic would be, the team initially had the idea to search for an outer space-related database (such as NASA). However, we only found options that we had concerns such as being too small, too clean, or not easily understood by non-experts of the subject. Eventually, we found common ground in the food area. The database of interest in this project is the Spoonacular Nutrition, Recipe, and Food API, which contains 365000 recipes and 86000 food products. This database boasts nutritional information, recipe costs, ingredient lists, equipment needs, special diets, diverse cuisines, and much more information a home cook would find useful. The potential of this API is endless but is of particular interest to this workgroup due to our shared interest in living healthy and nutrition.

The only change was regarding the paywall. My groupmate and I pooled resources to maximize our dataset, each making about 400 requests for a total of roughly 650 clean records. We then shared the compiled CSV data and continued separately from that point on.

Scrub

Once I had the full 650 “clean” csv records I ran them through python and realized that they were much worse than I had expected.

Question 1: First of all the equipment list was read as a string replicating a list format. I had to use the split, and replace functions to grab the relevant data and format it as a proper list. The most logical way to approach storing the information was a dictionary with recipeID as the key, and a tuple with spoon score and equipment count as the values. Unfortunately this process threw errors when trying to perform the cleaning on recipes with no equipment, so I introduced a if statement to handle recipes like this.

Once I had the equipment dictionary complete I could make a scatter plot of the equipment count and spoon score. This part was easy, however I thought the visual would be improved with a line of best fit. For the line of best fit I had to pass the spoon score / equipment length tuple through a numpy array, and then use the numpy polyfit function to generate the best fit equation.

Question 2: This one was a bit messy because of the huge scatter matrix. I think the scatter matrix has a tendency to dominate the page, but it's a really important visual to create whenever you are looking for correlations. Before the correlation matrix was even created I had to drop all quantitative fields (summary, title, links, etc.). This still left me with the vegan / vegetarian boolean fields, which can't be analysed directly. I used the get_dummies function of python to convert these boolean fields into two binary columns each, which can then be analysed in the same way as the other values.

Once the correlation was analysed and the regression factors selected, I had to remove the outliers from each column. I used the stats function from the scipy package to remove all values greater than three standard deviations away from the column average. Once the coefficients were generated I was able to plot the values without issue.

Question 3: This question was a bit simpler as most of the data had been cleaned answering the previous two questions. I used the same methodology to drop the price per serving outliers, and split the data frames into vegan and non-vegan data frames to plot separately. I did make sure to make both histogram plots equal axis to ensure transparency in the analysis.

Explore

To prevent statistical issues that arise when "data mining" I tried to stick close to my research questions, however I did do some exploration in question 2.

I noticed in the scatter matrix that there was a small linear correlation between some of the factors (ex. Spoon score and price per serving), but looking at the price per serving histogram the distribution was very skewed. I decided a log-transformed model would be best to fit this model. I added log transformation versions for the ready in minutes field as well to ensure the lack of correlation was not due to the skewed distribution.

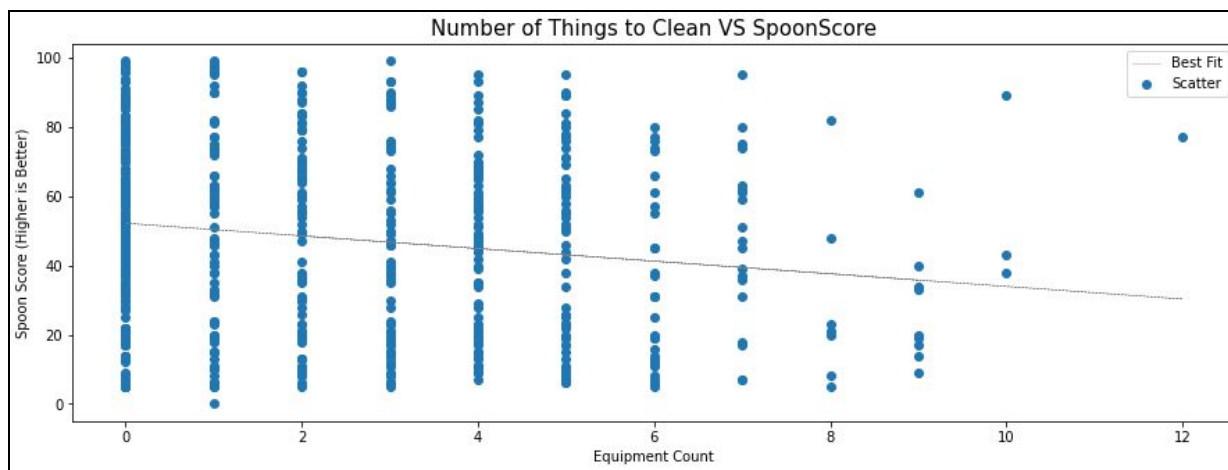
Apart from that I did try and push my skill-set by exploring matplotlib's visualization tools and creating great visuals.

Modeling

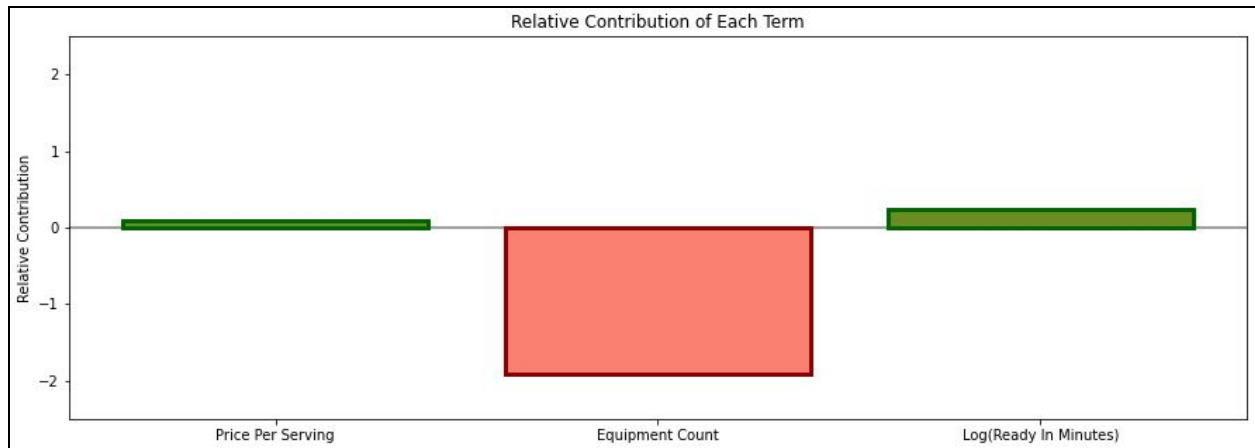
As discussed in the scrubbing section, all my questions analysed only data relevant to the question itself. This reduced dimensionality is apparent mostly in question 2 where I had to reduce the parameters to allow for a clean model fit.

Interpreting

Question 1: Surprisingly, the more equipment you bring out, the worse the spoon score gets. One would expect improvements in the food quality equivalent to the increase in work / cleanup, but in this case there is roughly a 1.5 point decline in score for each additional equipment item used in the recipe. Maybe that means the reviewers assigning the spoon score are lazy?

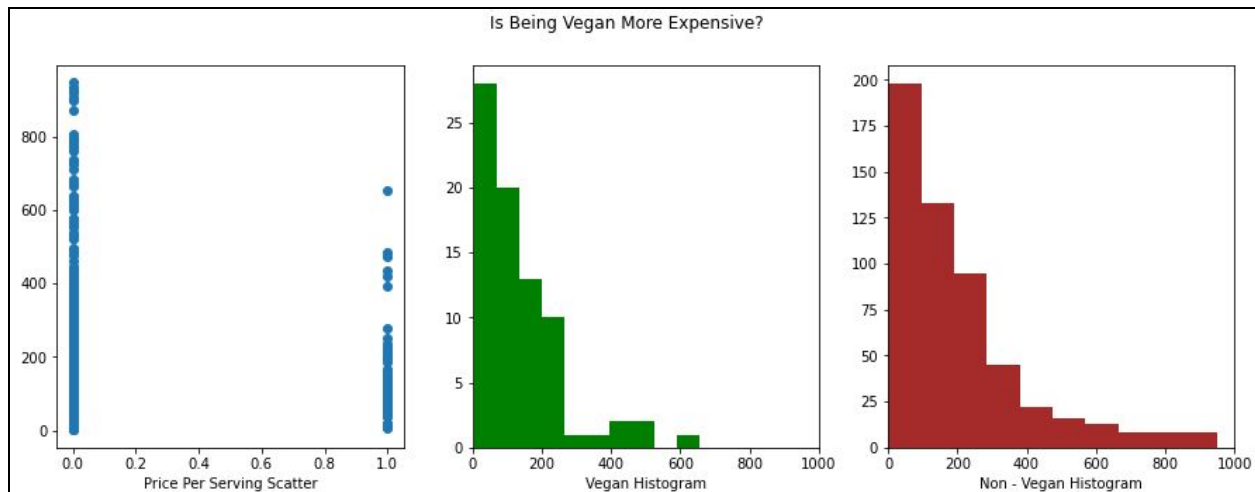


Question 2: It seems a bit odd that so many lines of code would produce such a simple visual but here it is. This shows the relative contribution of each relevant term to the model predicting spoon score. It's obvious from this model that as equipment count increases, spoon score decreases. It's less clear, but still statistically significant from additional R analysis, (*pvalues for intercept, priceperserving, equipmentlen, and logreadyinminutes are 1.4e-06, 2.6e-12, 0.03, and 0.04 respectively*), that price per serving and log(ready in minutes) positively correlate with spoon score.



Question 3: At first glance it looks like being vegan is less expensive than recipes with animal product, but that may not tell the whole story. The visual below shows a similar distribution from 0 to 200 price units, and then the vegan histogram runs out of data. Looking at the difference in counts, I believe it's just a coincidence that the vegan recipes are less expensive, and I don't think we have a representative data set here.

Another possible option is the lack of representation of vegan recipes in the spoon score API. I took a second look at the data and found that the average *ready in minutes* value for the vegan recipes is 67.4 minutes while the average non-vegan recipe took 88.43 minutes. I think this is another indicator that there are not many "full meal" vegan recipes, and therefore the dataset is not representative, and my question cannot be answered fully.



Points: total = 6 points

There was quite a bit of cleaning involved in our API request, which would give us **3 points** in the **not standardized format** category. As discussed in the scrub section, many of the fields were either missing, or incorrect which required additional wrangling. Each recipe is a unique request and processing requirement and is therefore **split across multiple files** for another **1 point**. The data also has **strings with punctuation** in the summary field, and is **connected via an API** for **one point each (6 total)**.

- **Data is not provided in a standardized format: 3pts**
 - For example, the MSDS files had some common elements but they were far from standardized, the LD 50 was expressed in many different ways
- **Data is split up across multiple files to begin with: 1pt**
- **Data is in a format other than one of the following: 2pts**
 - CSV
 - JSON
 - In a database
- **Data contains strings with punctuation (quotes or commas): 1pt**
- **Data set is larger than 1GB in size: 1pt**
- **Data set is composed of more than one type of related data: 2pts**
 - For example there is a table of products and a table of orders
- **Data set needs to be accessed in a way other than connecting to a database or downloading a file: 1pt**
 - I had to mount a virtual hard drive to get access to the MSDS files
 - You might need to make HTTP requests to an API

Table 1. Dataset Overview, Chosen fields of interest from recipe API

Title of field	Data type	Additional cleaning/data wrangling
Recipe id	int	Capture random recipe IDs
Recipe type	string	none
cuisine	string	none
Dish type	string	none
spoonscore	int	none
vegan	t/f	none
vegetarian	t/f	none
image URL	string (.jpg link)	none
title	string	none
source URL	string (site link)	none
Ready in minutes	int	none
Price per serving	float	none
summary	str	Remove html tags, capture first 200 chars
equipment	str	Pull equipment from recipe instructions. Do not include duplicate equipment

Sample Records:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	cuisines	dishTypes	spoonSco	Vegan	Vegetaria	ImageURL	Title	SourceUR	readyInM	PricePerS	Summary	equipment				
2	247533	['']	['side dish']	20	FALSE	TRUE	https://sp	Blueberry-Lemon Coffee Cake	http://ww	45	50.82	Blueberry ['oven', 'measuring cup', 'knife', 'whisk', 'bowl', 'ble					
3	240069	['']	['side dish']	59	FALSE	FALSE	https://sp	Falafel-Stuffed Pitas	http://ww	45	91.79	Falafel-Sti ['food processor', 'frying pan', 'whisk']					
4	255341	['']	['side dish']	41	FALSE	FALSE	https://sp	Stuffed Summer Squash di Alici	http://ww	60	126.62	Stuffed Su ['oven', 'frying pan']					
5	277083	['']	['side dish']	19	FALSE	TRUE	https://sp	Baked Brie with Honey	http://ww	45	54.51	Baked Brii ['oven', 'baking pan']					
6	124190	['English',	['side dish']	30	FALSE	TRUE	https://sp	Dried Cherry Scones	http://ww	60	89.86	Dried Cherry Scones might be just the morn meal you are sea					
7	27162	['']	['']	35	TRUE	TRUE	https://sp	Bread and Tomato Soup	http://ww	30	135.76	Bread and ['sauce pan']					
8	252211	['Mediterr	['lunch', 'n	48	FALSE	FALSE	https://sp	Risotto alla Milanese	http://allr	40	355.44	Risotto all ['slotted spoon', 'sauce pan', 'frying pan', 'ladle']					
9	145072	['']	['lunch', 'n	62	FALSE	FALSE	https://sp	Chipotle Lime Marinated Grilled Pork Chops or Tenderloin	http://ww	20	184.33	Chipotle Lime Marinated Grilled Pork Chops or Tenderloin is					
10	225936	['']	['lunch', 'n	89	FALSE	FALSE	https://sp	Broiled Salmon with Lemon and Olive Oil	http://ww	32	444.81	Broiled Sa ['ziploc bags', 'aluminum foil', 'broiler pan', 'whisk']					
11	46756	['']	['lunch', 'n	68	FALSE	FALSE	https://sp	Salmon and Dill Chowder	http://ww	45	329.58	Need a gli ['paper towels', 'sauce pan']					
12	25715	['Mexican	['side dish']	52	FALSE	TRUE	https://sp	Summer Squash Tacos With Sweet Corn & Queso Fresco	http://ww	45	179.54	Summer Squash Tacos With Sweet Corn & Queso Fresco migh					
13	236366	['']	['lunch', 'n	89	FALSE	FALSE	https://sp	Grilled Tuna with Rain Forest Glaze	http://ww	45	624.86	Grilled Tu ['sauce pan', 'whisk', 'bowl', 'frying pan', 'grill']					
14	203144	['']	['lunch', 'n	52	FALSE	FALSE	https://sp	Peasant Stew	http://ww	45	135.44	Peasant Si ['slow cooker', 'bowl', 'ladle']					
15	251831	['']	['side dish']	85	TRUE	TRUE	https://sp	Cheap and Easy Lentil Salad	http://allr	90	57.48	Cheap anc ['sauce pan', 'bowl', 'whisk', 'plastic wrap']					
16	179490	['']	['antipasti	8	FALSE	FALSE	https://sp	Black and White Pretzel Fudge	http://ww	70	13.03	Black and ['aluminum foil', 'frying pan', 'sauce pan', 'microwa					
17	56610	['']	['side dish']	20	FALSE	TRUE	https://sp	Sticky Toffee Cakelets	http://ww	2	97.02	Sticky Toffee Cakelets might be just the side dish you are sea					
18	93547	['Mexican	['lunch', 'n	73	FALSE	FALSE	https://sp	Simple Potato Tacos (Vegetarian)	http://ww	30	200.79	Simple Potato Tacos (Vegetarian) might be just the Mexican r					
19	254748	['America	['side dish']	99	TRUE	TRUE	https://sp	Slow Cooker Veggie Chili	http://allr	270	85.52	The recipe ['slow cooker']					
20	180364	['Mediterr	['side dish']	32	FALSE	FALSE	https://sp	Easy Italian Marinated Shrimp	http://ww	135	138.21	Easy Italia ['bowl']					
21	12946	['']	['antipasti	33	FALSE	TRUE	https://sp	Dixie Caviar Cups	http://ww	15	121.12	Dixie Cavi ['bowl']					
22	238857	['Creole',	['lunch', 'n	47	FALSE	FALSE	https://sp	Sausage Jambalaya	http://ww	45	225.06	Sausage Ji ['slow cooker']					
23	98668	['America	['lunch', 'n	75	FALSE	FALSE	https://sp	Wisconsin Badger Burger	http://ww	25	340.23	You can never have too many main course recipes, so give W					
24	59856	['']	['side dish']	40	FALSE	FALSE	https://sp	Caramel Walnut Tart	http://not	36	155.18	Caramel Walnut Tart might be just the dessert you are search					
25	158783	['']	['lunch', 'n	41	FALSE	FALSE	https://sp	Grilled Shrimp and Sausage Kabobs	http://ww	20	214.91	Grilled Sh ['metal skewers', 'grill', 'bowl']					
26	133585	['Mediterr	['beverage	5	FALSE	FALSE	https://sp	Tiramisu-Tini (The Best Tiramisu Martini)	http://ww	3	178.4	Tiramisu-Tini (The Best Tiramisu Martini) might be a good rec					
27	92153	['']	['side dish']	78	FALSE	TRUE	https://sp	Hot & Sour Soup-- 0 Points	http://he	30	120.3	Hot & Sour Soup-- 0 Points might be a good recipe to expand					
28	268695	['']	['antipasti	10	FALSE	FALSE	https://sp	Western Hash	http://ww	40	17.28	Western H ['frying pan']					
29	118547	['']	['lunch', 'n	55	FALSE	FALSE	https://sp	Roast Beef and Gravy	http://allr	490	204.68	Need a dairy free sauce? Roast Beef and Gravy could be an ar					
30	239692	['']	['side dish']	92	TRUE	TRUE	https://sp	Sweet Potato with Toasted Coconut	http://ww	30	61.36	Need a gli ['sauce pan', 'frying pan']					

An overview of the data wrangling process written out in words with code snippets where appropriate:

- This is discussed in the **Scrubbing** section of the **OSEMN** process

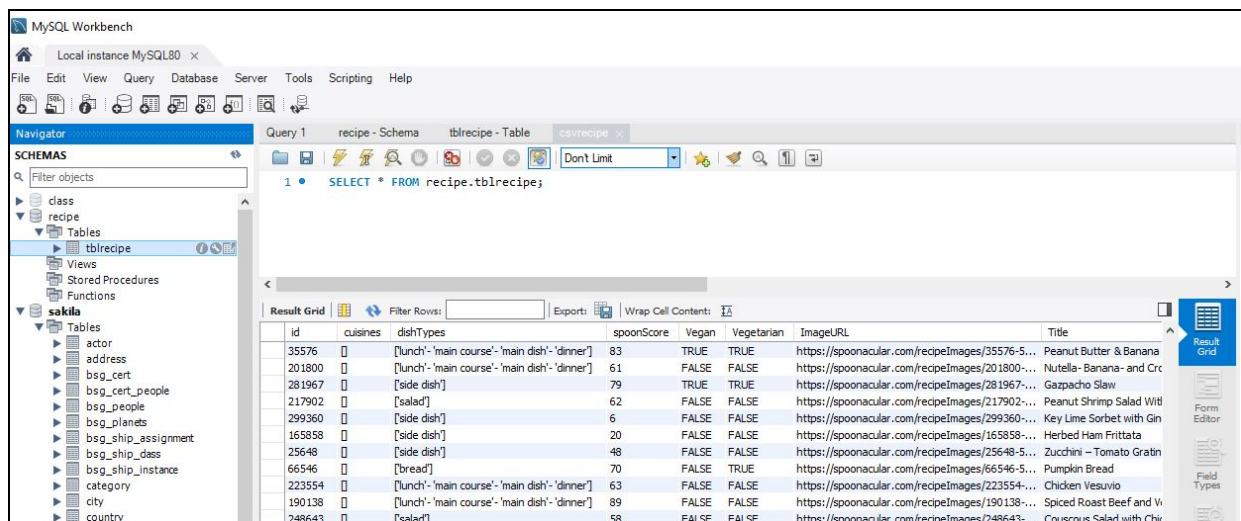
Database:

The data import process in MySQL Workbench is fairly straightforward, but there were a few hiccups. My strings in the CSV had some embedded commas, and therefore were unable to be read by the MySQL compiler. I had to remove all commas, and then the upload process could proceed. The data is flat and has no need to be broken up into relational tables, therefore the schema is only the tblrecipe table.

I chose id to be int data type, and the primary key. All other columns to be string data types excluding spoonScore, readyInMinutes, and PricePerServing which were int, int, and double respectively.

Process:

1. Download MySQLWorkbench
2. Create local instance of a database
3. Create Recipe schema using the embedded “Create Schema” wizard
4. Remove embedded comments from the CSV
5. Import CSV using the “import data” wizard



The screenshot shows the MySQL Workbench interface. The left sidebar displays the 'SCHEMAS' tree with 'recipe' selected. The main window shows a query result for 'SELECT * FROM recipe.tblrecipe;'. The result grid contains 15 rows of data with columns: id, cuisines, dishTypes, spoonScore, Vegan, Vegetarian, ImageURL, and Title.

id	cuisines	dishTypes	spoonScore	Vegan	Vegetarian	ImageURL	Title
35576		['lunch'-'main course'-'main dish'-'dinner']	83	TRUE	TRUE	https://spoonacular.com/recipeImages/35576-5...	Peanut Butter & Banana
201800		['lunch'-'main course'-'main dish'-'dinner']	61	FALSE	FALSE	https://spoonacular.com/recipeImages/201800-...	Nutella-Banana- and Crc
281967		['side dish']	79	TRUE	TRUE	https://spoonacular.com/recipeImages/281967-...	Gazpacho Slaw
217902		['salad']	62	FALSE	FALSE	https://spoonacular.com/recipeImages/217902-...	Peanut Shrimp Salad Wit
299360		['side dish']	6	FALSE	FALSE	https://spoonacular.com/recipeImages/299360-...	Key Lime Sorbet with Gn
165858		['side dish']	20	FALSE	FALSE	https://spoonacular.com/recipeImages/165858-...	Herbed Ham Frittata
25648		['side dish']	48	FALSE	FALSE	https://spoonacular.com/recipeImages/25648-5...	Zucchini - Tomato Gratin
66546		['bread']	70	FALSE	TRUE	https://spoonacular.com/recipeImages/66546-5...	Pumpkin Bread
223554		['lunch'-'main course'-'main dish'-'dinner']	63	FALSE	FALSE	https://spoonacular.com/recipeImages/223554-...	Chicken Vesuvio
190138		['lunch'-'main course'-'main dish'-'dinner']	89	FALSE	FALSE	https://spoonacular.com/recipeImages/190138-...	Spiced Roast Beef and V
248643		['salad']	58	FALSE	FALSE	https://spoonacular.com/recipeImages/248643-...	Couscous Salad with Chi