Tashina Taylor

1/25/2021

OSEMN Report

Problem Statement: Find data and convert it into a more usable form using Python, then explore the data in order to decide what questions to answer.

1. OBTAIN

   The dataset I chose for this project is Homebrew Beer Recipes for over 180,000 homemade beers. This dataset was obtained from Kaggle.com, from the user "Matteo" who scraped the data from 'brewersfriend.com'. All data was from before July 2020 and has not been updated since. The database is structured in JSON format and does not appear to have been cleaned or standardized by the user who uploaded it.

2. SCRUB

   While the data came in JSON format, it was not standardized. Some of the data came with # comments. The user who scraped the data and made the comments also does not appear to be a native English speaker, so trying to discern some of the comments was challenging.

   Creating the scripts to transform the data from JSON to CSV and back again was easy enough using pandas, but one issue that I ran into was that the data was being loaded wide instead of long. This was causing a "data not loaded completely" error at the startup of Excel opening the document, and only showed ~16,000 of the data points. Not only that, but it's quite hard to read.

   The solution to this issue was in the pandas documentation. All I had to do was add "orient='index'" to the df.reader in order to have the data sorted longwise and have the headers where they're supposed to go.

   Another issue was that the recipes should be a list, but came out as a single string per column. In order to bypass this (also found from the pandas documentation) was adding a "converters" function with the name of the header + "eval", which had pandas evaluate the data included and stop it from being a string when converted to JSON again. I could have further extracted these columns for additional data but chose not to do that for this point in the project just yet. I haven't decided if I want to use all that information or not.

3. EXPLORE

   The data consists of name, url, method, style, batch, og, fg, abv, ibu, color, ph mash, fermentables, hops, hops Summary, other, yeast, rating num, rating, and views.

   The basic data points are:

   a. "name" of course refers to the name of the beer being made.
   b. "url" links to the page of the beer on brewersfriend.com, but I will not be using this information.
   c. "method" refers to "all grain", "BIAB", "extract", or "partial mash".
      i. All grain brewing is the traditional method of making beer and used by just about all professional breweries. The brewer takes crushed malted grains and mashes them to convert starches into fermentable sugar.
      ii. BIAB is an acronym for "brewing beer in a bag" and is a different way of doing an "all grain" method, typically used by non-commercial backyard brewers.
      iii. "extract" In this process, the "all grain" work has already been done and the sugars are concentrated into a syrup or dry powder format.
      iv. "partial mash" - Essentially using a portion of the fermentable sugars for the wort from a mix of base and specialty grains.
   d. "style" is referring to the kind of beer being made, for example "lager", "IPA", "American Amber Ale", ect.
   e. "batch" is referring to how many gallons per batch (fermenter volume).
   f. "og" is for original gravity.
   g. "fg" is for final gravity.
   h. "abv" is for alcohol by volume.
   i. "ibu" is measured in "tinseth" and is basically a rating for how bitter a beer is (think IPA for high IBU).

j. <u>“ph mash”</u> is of course the pH of the mash – it should be noted that for this particular measurement in the original data, any N/A’s are “-1”.

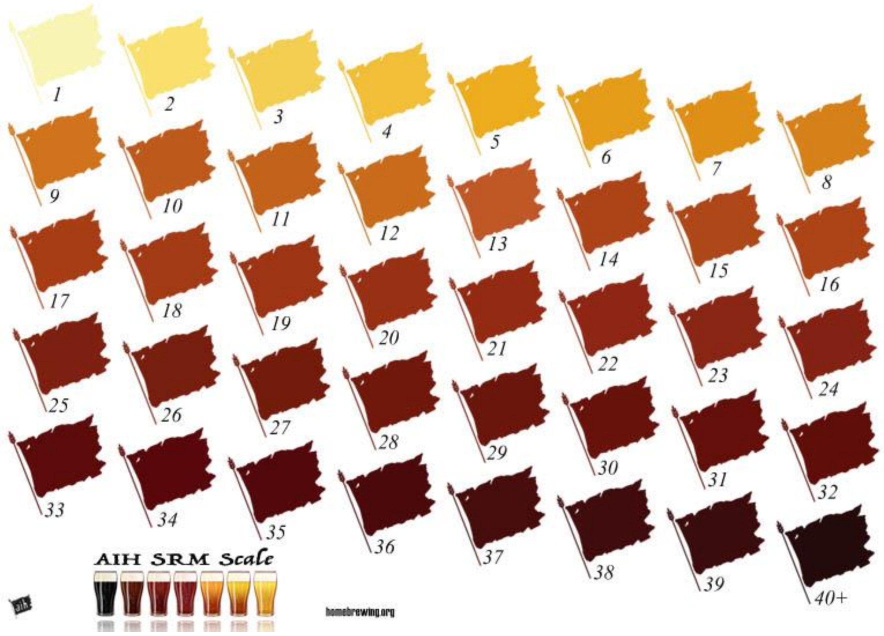
k. <u>“rating” </u>is based on a 5-star scale.

l. <u>“num rating”</u> is the number of how many reviews the brew has received to receive the“rating” number.

m. <u>“views”</u> is how many times the recipe page has been viewed at the time of the data collection. This also appears to be how the data was originally sorted – by page views rather than alphabetically.

The more complicated data:

n. <u>Color: </u> The number is based on the “Official AIH Standard Research method Number Scale”. The color of beer is determined by the grains and extracts that are used to make the brew. Attached is the basic color chart for reference. When looking at the data, there are some numbers that go far above the 40 range. There are several numbering methods for color of beer, and some of the data has extreme outliers that are probably typos. The assigned number should fall between one of these associated colors:



As mentioned above, there are some columns that contain multiple strings. These are the “fermentables”, “hops”, “hops summary”, “other”, and “yeast” columns.

o. <u>Fermentables:</u> This is a list of strings and integers that tells you the weight in kg (converted from lbs by the original scraper); the name of the grain that is used; PPG (this is the specific gravity you would get if you dissolved or mashed one pound of the ingredient in one gallon of water); °L (degrees Lintner is a unit used to measure the ability of a malt to reduce starch to sugar); and Bill% (called “grain bill” – this makes up what % each grain is used in the entirety of the fermentable recipe) – all bill % should sum up to 100%.

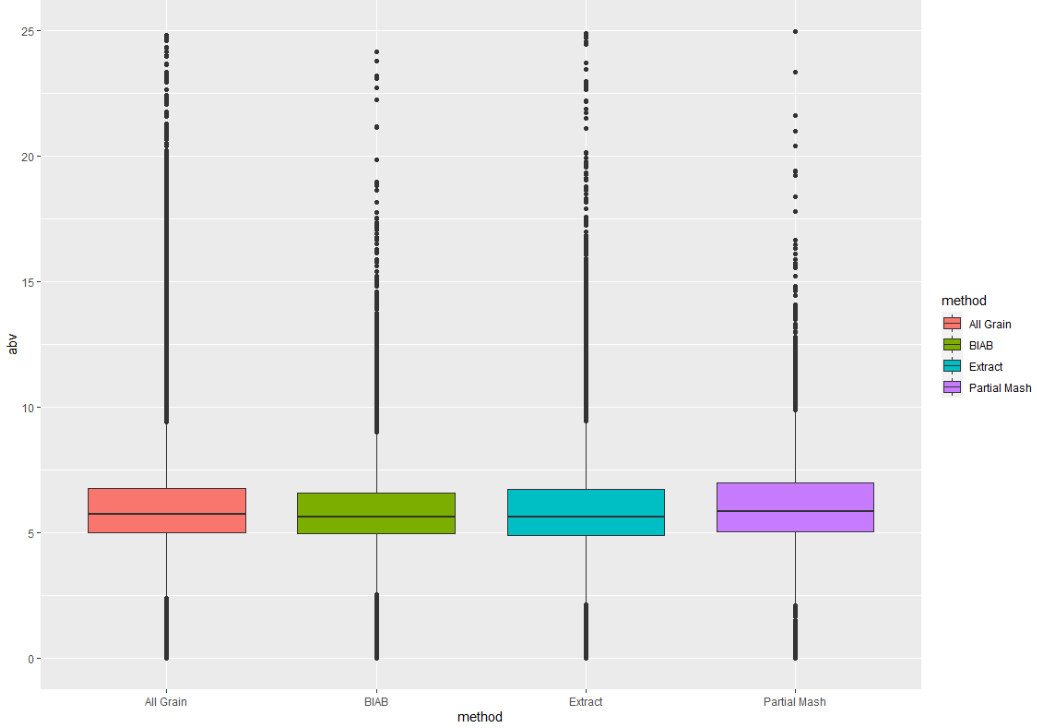
p. <u>Hops:</u> This of course is the list that tells about which hops to use. It starts with the weight of the hops in grams, converted from ounces; the name of the hops; how the hops have been processed (whole, pellet, leafs, ect); AA% (alpha acids – this contributes to the bitterness in beer. During the boil alpha acids are isomerized and increase IBUs); use type, like boiling, dry hop, whirlpool, ect; how long to process in minutes; IBU again, but related to hops (One Bitterness Unit is equal to 1 milligram of isomerized alpha acid in 1 liter of wort or beer or 1 part per million isomerized alpha acid. This is a system of measuring bitterness devised by brewing scientists and is an accepted standard throughout the world); and Bill %.

q. <u>Hops Summary:</u> The comment from the person who scraped the data says this is the “*sum of all the variety of hops disregarding when added*”. I am not quite sure what they mean by this, but I will probably not use this information for now.

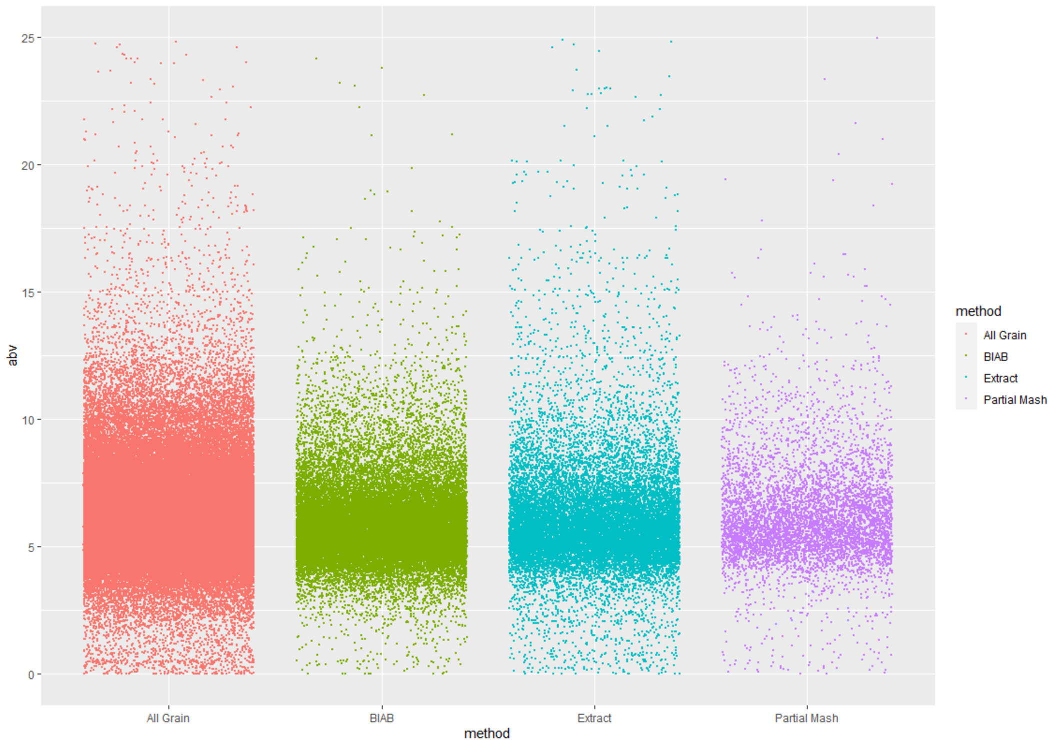
r. <u>“Other”:</u> This is for other additives in the recipe like vanilla extract. However, the comment from the person who scraped the data says “*other addition not much processing of the data done here (too variable)*” I will probably not use this information either.

s. “Yeast”: This tells us what kind of yeast was used for the fermentation process. It has the name of the yeast; yeast attenuation (the degree to which yeast ferments the sugar in a wort or must. If you have 50% attenuation it means that 50% of the sugars have been converted into alcohol and CO2 by yeast.); yeast flocculation (yeast coming together and dropping to the bottom of a fermenter) and measured in “high” or “low”; low and then high optimum temperatures in degrees Fahrenheit; and whether or not a starter was used (A starter is simply a small volume of wort that's used for the sole purpose of growing yeast cells).

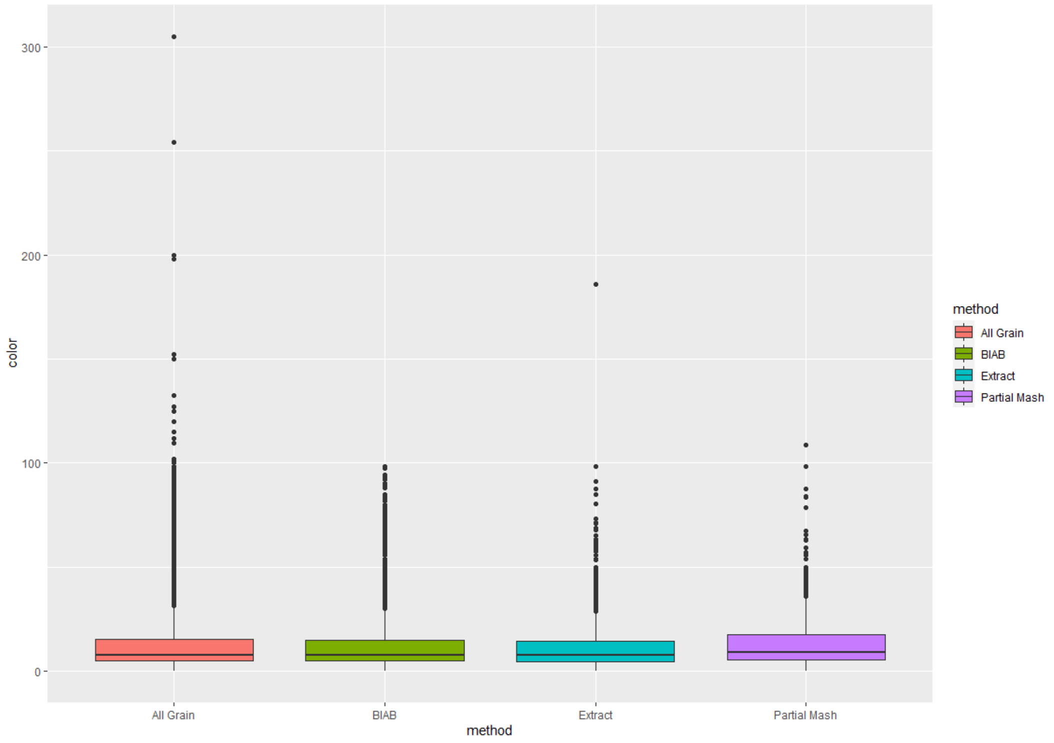
Below I have included a few initial exploratory plots and facts:
- the mean ABV of beer is 6.02%
- the mean IBU is 38.90
- the mean color is 12.81 (red/orange color)



This is a boxplot showing methods vs. abv



This is a jitterplot showing method vs abv



Here is a boxplot of color vs. method. I may need to normalize the color data...

4. MODEL
   Depending on what data I want to compare, there are a few model options. Initially, I am going to do a One-way ANOVA test on the brewing methods vs ABV to see if there is any difference between the 4 groups of methods (All grain, BIAB, Extract, Partial Mash). From there I would like to do the same vs other popular parameters like IBU and color.

5. INTERPRET
   Data has only been lightly explored at this point, and I still need to decide on how to further wrangle it, probably with relational databases. Interpretation would not apply as nothing has been modeled.

Initial Data Sample(JSON format in text file):

{"0": {"name": "Vanilla Cream Ale", "url": "/homebrew/recipe/view/1633/vanilla-cream-ale", "method": "All Grain", "style": "Cream Ale", "batch": 21.8, "og": 1.055, "fg": 1.013, "abv": 5.48, "ibu": 19.44, "color": 4.83, "ph mash": -1, "fermentables": [[2.381, "American - Pale 2-Row", 37.0, 1.8, 44.7], [0.907, "American - White Wheat", 40.0, 2.8, 17.0], [0.907, "American - Pale 6-Row", 35.0, 1.8, 17.0], [0.227, "Flaked Corn", 40.0, 0.5, 4.3], [0.227, "American - Caramel / Crystal 20L", 35.0, 20.0, 4.3], [0.227, "American - Carapils (Dextrine Malt)", 33.0, 1.8, 4.3], [0.113, "Flaked Barley", 32.0, 2.2, 2.1], [0.34, "Honey", 42.0, 2.0, 6.4]], "hops": [[14.0, "Cascade", "Pellet", 6.2, "Boil", "60 min", 11.42, 33.3], [14.0, "Cascade", "Pellet", 6.2, "Boil", "20 min", 6.92, 33.3], [14.0, "saaz", "Pellet", 3.0, "Boil", "5 min", 1.1, 33.3]], "hops Summary": [[28.0, "Cascade (Pellet)", 18.34, 66.6], [14.0, "saaz (Pellet)", 1.1, 33.3]], "other": [["2 oz", "pure vanilla extract", "Flavor", "Boil", "0 min."], ["1 oz", "pure vanilla extract", "Flavor", "Bottling", "0 min."], ["1 tsp", "yeast nutrient", "Other", "Boil", "15 min."], ["1 each", "whirlfloc", "Fining", "Boil", "15 min."], ["4 each", "Vanilla beans - in 2oz Vodka", "Other", "Secondary", "0 min."]], "yeast": ["Wyeast - K\u00f6lsch 2565", "76%", "Low", "56", "70", "Yes"], "rating": 0, "num rating": 16, "views": 289454},
"1": {"name": "Avg. Perfect Northeast IPA (NEIPA)", "url": "/homebrew/recipe/view/363082/avg-perfect-northeast-ipa-neipa-", "method": "All Grain", "style": "Specialty IPA: New England IPA", "batch": 21.8, "og": 1.062, "fg": 1.013, "abv": 6.5, "ibu": 59.26, "color": 5.2, "ph mash": 5.49, "fermentables": [[4.876, "American - Pale 2-Row", 37.0, 1.8, 77.0], [0.635, "American - Wheat", 38.0, 1.8, 10.0], [0.635, "Flaked Oats", 33.0, 2.2, 10.0], [0.191, "Canadian - Honey Malt", 37.0, 25.0, 3.0]], "hops": [[28.0, "Citra", "Pellet", 12.6, "Boil", "10 min", 13.55, 7.3], [28.0, "Galaxy", "Pellet", 15.6, "Boil", "10 min", 16.77, 7.3], [42.0, "Citra", "Pellet", 12.6, "Whirlpool at 170 \u00b0F", "15 min", 9.23, 10.9], [42.0, "Galaxy", "Pellet", 15.6, "Whirlpool at 170 \u00b0F", "15 min", 11.43, 10.9], [42.0, "Mosaic", "Pellet", 11.3, "Whirlpool at 170 \u00b0F", "15 min", 8.28, 10.9], [28.0, "Citra", "Pellet", 12.6, "Dry Hop", "7 days", 0, 7.3], [42.0, "Galaxy", "Pellet", 15.6, "Dry Hop", "7 days", 0, 10.9], [28.0, "Mosaic", "Pellet", 11.3, "Dry Hop", "7 days", 0, 7.3], [35.0, "Citra", "Pellet", 12.6, "Dry Hop", "3 days", 0, 9.1], [42.0, "Galaxy", "Pellet", 15.6, "Dry Hop", "3 days", 0, 10.9], [28.0, "Mosaic", "Pellet", 11.3, "Dry Hop", "3 days", 0, 7.3]], "hops Summary": [[133.0, "Citra (Pellet)", 22.78, 34.6], [154.0, "Galaxy (Pellet)", 28.2, 40.0], [98.0, "Mosaic (Pellet)", 8.28, 25.5]], "other": [["0.50 tsp", "Irish Moss", "Fining", "Boil", "15 min."], ["0.50 tsp", "Yeast Nutrient", "Other", "Boil", "15 min."], ["4 g", "Calcium Chloride", "Water Agt", "Mash", "1 hr."], ["5 g", "Gypsum", "Water Agt", "Mash", "1 hr."], ["6 ml", "Phosphoric acid", "Water Agt", "Mash", "1 hr."]], "yeast": ["Wyeast - London Ale III 1318", "78%", "High", "64", "74", "No"], "rating": 0, "num rating": 23, "views": 288318},
"2": {"name": "Sierra Nevada Pale Ale Clone", "url": "/homebrew/recipe/view/28546/sierra-nevada-pale-ale-clone", "method": "All Grain", "style": "American Pale Ale", "batch": 24.6, "og": 1.055, "fg": 1.013, "abv": 5.58, "ibu": 39.79, "color": 8.0, "ph mash": 5.67, "fermentables": [[5.216, "American - Pale 2-Row", 37.0, 1.8, 92.7], [0.412, "American - Caramel / Crystal 60L", 34.0, 60.0, 7.3]], "hops": [[14.0, "Magnum", "Pellet", 15.0, "Boil", "60 min", 22.62, 8.3], [14.0, "Perle", "Pellet", 8.2, "Boil", "30 min", 9.51, 8.3], [28.0, "Cascade", "Pellet", 7.0, "Boil", "10 min", 7.66, 16.7], [56.0, "Cascade", "Pellet", 7.0, "Boil", "0 min", 0, 33.3], [56.0, "Cascade", "Pellet", 7.0, "Dry Hop", "4 days", 0, 33.3]], "hops Summary": [[14.0, "Magnum (Pellet)", 22.62, 8.3], [14.0, "Perle (Pellet)", 9.51, 8.3], [140.0, "Cascade (Pellet)", 7.66, 83.3]], "other": [["1 each", "Crush whilrfoc Tablet", "Water Agt", "Boil", "10 min."]], "yeast": ["Fermentis - Safale - American Ale Yeast US-05", "76%", "Medium", "54", "77", "Yes"], "rating": 0, "num rating": 26, "views": 271945}

CSV data (after converting it from the JSON format in the text file):

```
0	Vanilla Cream Ale/homebrew/recipe/view/1633/vanilla-cream-ale	All Grain	Cream Ale	21.8	1.055
	1.013	5.48	19.44	4.83	-1	[[2.3810000000000002, 'American - Pale 2-Row', 37.0, 1.8, 44.7],
[0.907, 'American - White Wheat', 40.0, 2.8, 17.0], [0.907, 'American - Pale 6-Row', 35.0, 1.8, 17.0], [0.227, 'Flaked Corn',
40.0, 0.5, 4.3], [0.227, 'American - Caramel / Crystal 20L', 35.0, 20.0, 4.3], [0.227, 'American - Carapils (Dextrine Malt)', 33.0,
1.8, 4.3], [0.113, 'Flaked Barley', 32.0, 2.2, 2.1], [0.34, 'Honey', 42.0, 2.0, 6.4]]	[[14.0, 'Cascade', 'Pellet', 6.2, 'Boil', '60
min', 11.42, 33.3], [14.0, 'Cascade', 'Pellet', 6.2, 'Boil', '20 min', 6.92, 33.3], [14.0, 'saaz', 'Pellet', 3.0, 'Boil', '5 min', 1.1, 33.3]]
	[[28.0, 'Cascade (Pellet)', 18.34, 66.6], [14.0, 'saaz (Pellet)', 1.1, 33.3]]	[['2 oz', 'pure vanilla extract', 'Flavor', 'Boil',
'0 min.'], ['1 oz', 'pure vanilla extract', 'Flavor', 'Bottling', '0 min.'], ['1 tsp', 'yeast nutrient', 'Other', 'Boil', '15 min.'], ['1 each',
'whirlfloc', 'Fining', 'Boil', '15 min.'], ['4 each', 'Vanilla beans - in 2oz Vodka', 'Other', 'Secondary', '0 min.']]	['Wyeast - Kölsch
2565', '76%', 'Low', '56', '70', 'Yes']	0	16	289454
1	Avg. Perfect Northeast IPA (NEIPA)	/homebrew/recipe/view/363082/avg-perfect-northeast-ipa-neipa-	All Grain
	Specialty IPA: New England IPA	21.8	1.062	1.013	6.5	59.26	5.2	5.49	[[4.876,
'American - Pale 2-Row', 37.0, 1.8, 77.0], [0.635, 'American - Wheat', 38.0, 1.8, 10.0], [0.635, 'Flaked Oats', 33.0, 2.2, 10.0],
[0.191, 'Canadian - Honey Malt', 37.0, 25.0, 3.0]]	[[28.0, 'Citra', 'Pellet', 12.6, 'Boil', '10 min', 13.55, 7.3], [28.0, 'Galaxy',
'Pellet', 15.6, 'Boil', '10 min', 16.77, 7.3], [42.0, 'Citra', 'Pellet', 12.6, 'Whirlpool	at 170 °F', '15 min', 9.23, 10.9],
[42.0, 'Galaxy', 'Pellet', 15.6, 'Whirlpool	at 170 °F', '15 min', 11.43, 10.9], [42.0, 'Mosaic', 'Pellet', 11.3, 'Whirlpool
at 170 °F', '15 min', 8.28, 10.9], [28.0, 'Citra', 'Pellet', 12.6, 'Dry Hop', '7 days', 0, 7.3], [42.0, 'Galaxy', 'Pellet', 15.6, 'Dry Hop',
'7 days', 0, 10.9], [28.0, 'Mosaic', 'Pellet', 11.3, 'Dry Hop', '7 days', 0, 7.3], [35.0, 'Citra', 'Pellet', 12.6, 'Dry Hop', '3 days', 0,
9.1], [42.0, 'Galaxy', 'Pellet', 15.6, 'Dry Hop', '3 days', 0, 10.9], [28.0, 'Mosaic', 'Pellet', 11.3, 'Dry Hop', '3 days', 0, 7.3]]
	[[133.0, 'Citra (Pellet)', 22.78, 34.6], [154.0, 'Galaxy (Pellet)', 28.2, 40.0], [98.0, 'Mosaic (Pellet)', 8.28, 25.5]]
	[['0.50 tsp', 'Irish Moss', 'Fining', 'Boil', '15 min.'], ['0.50 tsp', 'Yeast Nutrient', 'Other', 'Boil', '15 min.'], ['4 g',
'Calcium Chloride', 'Water Agt', 'Mash', '1 hr.'], ['5 g', 'Gypsum', 'Water Agt', 'Mash', '1 hr.'], ['6 ml', 'Phosphoric acid', 'Water
Agt', 'Mash', '1 hr.']]	['Wyeast - London Ale III 1318', '78%', 'High', '64', '74', 'No']	0	23	288318
2	Sierra Nevada Pale Ale Clone	/homebrew/recipe/view/28546/sierra-nevada-pale-ale-clone	All Grain
	American Pale Ale	24.6	1.055	1.013	5.58	39.79	8	5.67	[[5.216, 'American - Pale
2-Row', 37.0, 1.8, 92.7], [0.41200000000000003, 'American - Caramel / Crystal 60L', 34.0, 60.0, 7.3]]	[[14.0,
'Magnum', 'Pellet', 15.0, 'Boil', '60 min', 22.62, 8.3], [14.0, 'Perle', 'Pellet', 8.2, 'Boil', '30 min', 9.51, 8.3], [28.0, 'Cascade',
'Pellet', 7.0, 'Boil', '10 min', 7.66, 16.7], [56.0, 'Cascade', 'Pellet', 7.0, 'Boil', '0 min', 0, 33.3], [56.0, 'Cascade', 'Pellet', 7.0, 'Dry
Hop', '4 days', 0, 33.3]]	[[14.0, 'Magnum (Pellet)', 22.62, 8.3], [14.0, 'Perle (Pellet)', 9.51, 8.3], [140.0, 'Cascade (Pellet)',
7.66, 83.3]]	[['1 each', 'Crush whilrfoc Tablet', 'Water Agt', 'Boil', '10 min.']]	['Fermentis - Safale - American Ale
Yeast US-05', '76%', 'Medium', '54', '77', 'Yes']	0	26	271945
```

JSON file data after converting it from CSV(with spacing = 4, and with 4 columns so it stays on 1 page):

```
{
    "0":{
        "Unnamed: 0":0,
        "name":"Vanilla Cream Ale",
        "url":"\/homebrew\/recipe\/view\/1633\/vanilla-cream-ale",
        "method":"All Grain",
        "style":"Cream Ale",
        "batch":21.8,
        "og":1.055,
        "fg":1.013,
        "abv":5.48,
        "ibu":19.44,
        "color":4.83,
        "ph mash":-1.0,
        "fermentables":[
            [
                2.381,
                "American - Pale 2-Row",
                37.0,
                1.8,
                44.7
            ],
            [
                0.907,
                "American - White Wheat",
                40.0,
                2.8,
                17.0
            ],
            [
                0.907,
                "American - Pale 6-Row",
                35.0,
                1.8,
                17.0
            ],
            [
                0.227,
                "Flaked Corn",
                40.0,
                0.5,
                4.3
            ],
            [
                0.227,
                "American - Caramel \/ Crystal 20L",
                35.0,
                20.0,
                4.3
            ],
            [
                0.227,
                "American - Carapils (Dextrine Malt)",
                33.0,
                1.8,
                4.3
            ],
            [
                0.113,
                "Flaked Barley",
                32.0,
                2.2,
                2.1
            ],
            [
                0.34,
                "Honey",
                42.0,
                2.0,
                6.4
            ]
        ],
        "hops":[
            [
                14.0,
                "Cascade",
                "Pellet",
                6.2,
                "Boil",
                "60 min",
                11.42,
                33.3
            ],
            [
                14.0,
                "Cascade",
                "Pellet",
                6.2,
                "Boil",
                "20 min",
                6.92,
                33.3
            ],
            [
                14.0,
                "saaz",
                "Pellet",
                3.0,
                "Boil",
                "5 min",
                1.1,
                33.3
            ]
        ],
        "hops Summary":[
            [
                28.0,
                "Cascade (Pellet)",
                18.34,
                66.6
            ],
            [
                14.0,
                "saaz (Pellet)",
                1.1,
                33.3
            ]
        ],
        "other":[
            [
                "2 oz",
                "pure vanilla extract",
                "Flavor",
                "Boil",
                "0 min."
            ],
            [
                "1 oz",
                "pure vanilla extract",
                "Flavor",
                "Bottling",
                "0 min."
            ],
            [
                "1 tsp",
                "yeast nutrient",
                "Other",
                "Boil",
                "15 min."
            ],
            [
                "1 each",
                "whirlfloc",
                "Fining",
                "Boil",
                "15 min."
            ],
            [
                "4 each",
                "Vanilla beans - in 2oz Vodka",
                "Other",
                "Secondary",
                "0 min."
            ]
        ],
        "yeast":[
            "Wyeast - K\u00f6lsch 2565",
            "76%",
            "Low",
            "56",
            "70",
            "Yes"
        ],
        "rating":0,
        "num rating":16,
        "views":289454
    },
```

JSON to CSV:

```
import pandas as pd


#opening JSON file and loading data
#orient='index' added to make sure data loaded long instead of wide
with open("recipes_full.txt", encoding='utf-8-sig') as beer:
    df = pd.read_json(beer, orient='index')


#write the converted csv file
df.to_csv('beer_list.csv', encoding='utf-8-sig', index=True)
```

CSV to JSON:

```
import pandas as pd


#read the csv
#converters is a pandas dict of functions for converting values in columns. In this case, I converted these values from a single string, to a list
df = pd.read_csv(r'beer_list.csv', converters={'fermentables':eval, 'hops':eval, 'hops Summary':eval, 'other':eval, 'yeast':eval})


#write the JSON file with an indent of 4
df.to_json (r'beer_list_json.json', orient="index", indent=4)
```