Visualization of data 2018

Course Project

Visualization in Recipes

Submitted by: ***Liad Nahmias***

User Tasks

Detect Sets and Exploration

What - Data Set

The dataset was collected by Kaggle1.

The dataset displays 20,052 different recipes with a specific rating between 0 to 5, calories, protein, fat, sodium and a list of ingredients as features (0 if this specific ingredient isn't in the recipe and 1 otherwise).

The data was required some cleaning actions:

- All the recipes with 0 ingredients
- All the definitions of the recipe – alcoholic, the destination country etc.,
- all the beverages and their ingredients

After all those cleaning actions, the ready data set was included 13,805 unique recipes with the features: (Recipes_Dataset_Full.csv)

- rating
- calories
- protein
- fat
- sodium
- list of ingredients that are in the recipe

ID	title	rating	calories	protein	fat	sodium	22-minutes	advance-p	breakfast	brunch	dairy	dessert	dinner	healthy	gluten-free	kid-friendly	kosher	kosher-pas	low-choles
1	Lentil Appl	2.5	426	30	7	559	0	0	0	0	0	0	0	0	0	1	0	0	0
2	Boudin Bla	4.375	403	18	23	1439	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Potato and	3.75	165	6	7	165	0	0	0	0	1	0	0	0	0	0	0	0	0
4	The Best E	4.375	948	19	79	1042	0	0	0	0	0	0	0	0	0	1	0	0	0
5	Korean Ma	4.375	170	7	10	1272	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Ham Persi	3.75	602	23	41	1696	0	0	0	0	0	0	0	0	1	0	0	0	0
7	Yams Brai	3.75	256	4	5	30	0	0	0	0	1	0	0	0	0	0	0	0	0
8	Banana-Cl	4.375	766	12	48	439	0	0	0	0	0	1	0	0	0	1	1	0	0
9	Beef Tend	4.375	174	11	12	176	0	0	0	0	0	0	0	0	1	0	0	0	0
10	Peach Mus	3.125	134	4	3	1394	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Raw Crear	4.375	382	5	31	977	0	0	0	0	0	0	0	1	0	0	0	0	0
12	Sweet Butt	1.875	146	4	5	160	0	0	1	1	0	1	0	0	1	0	1	0	0
13	Crisp Brais	4.375	890	59	68	1027	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Mozzarella	5	107	5	7	344	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Tuna Aspa	5	421	10	33	383	0	0	0	0	0	0	0	0	1	0	1	0	0
16	Asian Pear	4.375	345	11	19	423	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Sea Salt-R	3.75	279	3	30	206	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Garlic Bag	0	95	1	7	103	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Cucumber-	3.75	215	6	20	250	0	1	0	0	0	0	1	0	1	0	1	0	0
20	Dried Pear	2.5	14	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
21	Green Bea	4.375	351	6	19	79	0	0	0	0	0	0	0	0	1	0	1	0	0
22	Apricot-Ch	4.375	311	5	5	226	0	0	0	0	0	1	0	0	0	0	1	0	0
23	Roasted S	4.375	376	7	18	604	0	0	0	0	0	0	0	0	1	0	0	0	0
24	Deviled Ha	3.125	185	10	13	765	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Aztec Chic	3.75	625	39	44	1248	0	0	0	0	0	0	1	0	1	0	0	0	0
26	Sauteed B	4.375	107	4	10	329	0	0	0	0	0	0	0	0	1	0	1	0	0
27	Grouper w	4.375	336	44	16	413	0	0	0	0	0	0	0	0	0	0	0	0	0
28	Better-Tha	2.5	145	3	6	208	0	0	0	0	0	0	0	0	0	0	1	0	0

[1] https://www.kaggle.com/hugodarwood/epirecipes/data

Why – define the user tasks

After the cleaning actions, I saw in my eyes to main parts of the data – the recipes and the ingredients.

User tasks for each part:

1. Ingredients – **Detect Sets** as a function of ingredients connections and appearances in the context of the recipe dataset.
   Detecting sets, more specifically, pairs of ingredients will show us the inner correlations. Moreover, we will see in the "HOW" phase, that the correlation can be significant for only one of the ingredients in the pair or for the both equally.

2. Recipes – **Exploration** a specific recipe by sorting and filtering.
   When you get this amount of recipes, like looking in a very big recipes book, you want to find a specific recipe for cooking. You have specific requirement – dessert, lunch etc. and you want to choose between all the relevant recipes, the one that suits for you.

How – The implemented visualizations

I created two visualization, one for each user task:

1. Heat Map (index_heat_map.html)

   Part: Ingredients

   Data for visualization: (heat.csv)

   I manipulated the dataset with Python to create the required data for the map.

   I created a CSV file consists four columns: Left – ingredient name for the left column, Top - ingredient name for the top row, value – percentages (left ingredient and Top ingredient total number of pairs divided the left ingredient total appearances in recipes), Group – enter the values to 11 buckets ([0-4] to group 0, [5-14] to group 1 etc.).

   Thus, e.g. Left (arugula) and Top (basil) is different than Left (basil) and Top (arugula).

Left	Top	Value	Group
almond	almond	100	10
almond	anchovy	0	0
almond	anise	0	0
almond	apple	5	1
almond	apple-juice	0	0
almond	apricot	6	1
almond	artichoke	0	0
almond	arugula	0	0
almond	asian-pear	0	0
almond	asparagus	0	0
almond	avocado	0	0
almond	bacon	0	0
almond	banana	0	0
almond	barley	0	0
almond	basil	1	0
almond	bass	0	0
almond	bean	0	0
almond	beef	0	0
almond	beef-rib	0	0
almond	beef-shank	0	0
almond	beef-tende	0	0
almond	beer	0	0
almond	beet	0	0
almond	bell-peppe	2	0
almond	berry	6	1
almond	biscuit	0	0
almond	blackberry	1	0
almond	blue-chees	0	0
almond	blueberry	0	0

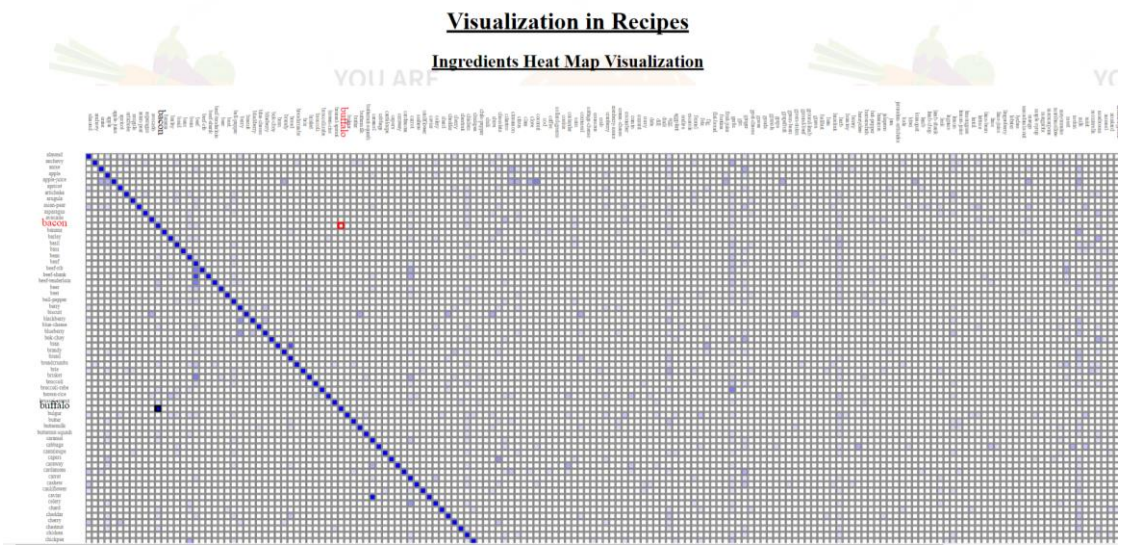
   The translate data to visualization aspects:

   - Every row and column are ingredients **(label)**
   - Every rectangle is the connection between two ingredients **(position)**
   - Rectangle fill shows the group of values by color hue (the more darker – more connections between two ingredients in the left ingredient recipes) **(color)**
   - The ingredients sorted by alphabetic order **(position)**

   Interaction:

- Hover every rectangle – shows the chosen rectangle with black border and paint the names of the both ingredients. In addition, to show comparison to the opposite tuple, it represents in red
- Zoom in and zoom out
- Dragging the visualization to focus in specific area
- Click on a square – will alert the values of the clicked black square and the opposite red square

The visualization:



Pros and Cons:

- Time:
  - + Very fast for getting insights
  - - Takes some time for every ingredient analysis (need to zoom in and check all the 279 other ingredients)
- Insights:
  - + Can show the most "interacted" ingredients (the darker colors)
  - + Detect different types of sets
- Essence:
  - + Can lead for the common ingredients you will need if you have specific ingredient at home.
  - + Because of the specific value calculation, good usage of all the table's cells

- Must using the interaction for zooming in.

- Confidence:

  + Uses all the relevant data and show it all in the same time

General speaking:

To sum up the pros and cons of this visualization to my project, the visualization shows the details of the connections between sets of two ingredients in the dataset and normalized it to the number of appearance of the specific ingredient in the dataset. Thus, the connection between to ingredients can be very high but not significant for one of them.

Although you can drag and zoom in and out, the area of the map is quite big and it hard to see the "big picture" in a glance.

Interesting Insights:

- The set "Buffalo" and "Bacon" is very significant for "Buffalo" (most of the recipes consists Buffalo also consists "Bacon") But it is not vice versa.
- The set "Marshmallow" and "Chocolate" is very significant for "Marshmallow" But it is not vice versa.
- The set "Flat-bread" and "Grape" is very significant for "Flat-bread" But it is not vice versa.
- We can see that Onion and tomato are highly connect to a lot of ingredients. Which means they are part of a lot of recipes.
- Sangria is highly connect to 7 ingredients – because it appears only in one recipe.
- Shellfish and shrimp are mutual connected – which means the both of them appears together in the same relative part of the recipes.

2. Matrix combined with Bars and Labels (index_matrix_combination.html)

Part: Recipes

Data for visualization: (Recipes_Dataset_Ready.csv, Ing_graph.json)

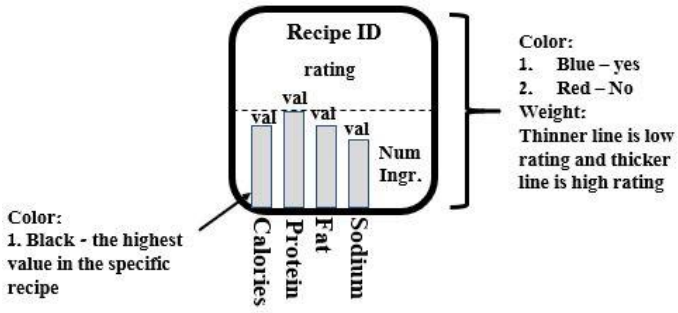
I manipulated the dataset with Python to create the required data for the combined matrix.

I created a CSV file consists those columns: ID – an ordered numbers, title – recipe name, rating – recipe's rating, calories – calories value, protein – protein value, fat – fat value, sodium – sodium value, binary traits (22- minutes, advance-prep, breakfast, brunch, dairy, dessert, dinner, healthy, gluten-free, kid-friendly, kosher, kosher-Passover, low-cholesterol, low-sugar, lunch, meat, paleo, pescatrian, seafood, vegan, vegetarian), ingredients – number of ingredients in the recipe.

ID	title	rating	calories	protein	fat	sodium	22-minutes	advance-p	breakfast	brunch	dairy	dessert	dinner	healthy	gluten-free	kid-friendly	kosher	kosher-pas	low-choles	low-sugar
1	Lentil Appl	2.5	426	30	7	559	0	0	0	0	0	0	0	0	0	1	0	0	0	0
2	Boudin Bla	4.375	403	18	23	1439	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Potato and	3.75	165	6	7	165	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	The Best E	4.375	948	19	79	1042	0	0	0	0	0	0	0	0	0	1	0	0	0	0
5	Korean Ma	4.375	170	7	10	1272	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Ham Persi	3.75	602	23	41	1696	0	0	0	0	0	0	0	0	1	0	0	0	0	0
7	Yams Brai	3.75	256	4	5	30	0	0	0	0	1	0	0	0	0	0	0	0	0	0
8	Banana-Cl	4.375	766	12	48	439	0	0	0	0	0	1	0	0	0	1	1	0	0	0
9	Beef Tend	4.375	174	11	12	176	0	0	0	0	0	0	0	0	1	0	0	0	0	0
10	Peach Mu	3.125	134	4	3	1394	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Raw Crear	4.375	382	5	31	977	0	0	0	0	0	0	0	1	0	0	0	0	0	0
12	Sweet But	1.875	146	4	5	160	0	0	1	1	0	1	0	0	1	0	1	0	0	0
13	Crisp Brais	4.375	890	59	68	1027	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Mozzarella	5	107	5	7	344	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Tuna Aspa	5	421	10	33	383	0	0	0	0	0	0	0	0	1	0	1	0	0	0
16	Asian Pear	4.375	345	11	19	423	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	Sea Salt-R	3.75	279	3	30	206	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Garlic Bag	0	95	1	7	103	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Cucumber	3.75	215	6	20	250	0	1	0	0	0	0	1	0	1	0	1	0	0	0
20	Dried Pear	2.5	14	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
21	Green Bea	4.375	351	6	19	79	0	0	0	0	0	0	0	0	1	0	1	0	0	0
22	Apricot-Ch	4.375	311	5	5	226	0	0	0	0	0	1	0	0	0	0	1	0	0	0
23	Roasted S	4.375	376	7	18	604	0	0	0	0	0	0	0	0	1	0	0	0	0	0
24	Deviled Ha	3.125	185	10	13	765	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Aztec Chic	3.75	625	39	44	1248	0	0	0	0	0	0	1	0	1	0	0	0	0	0
26	Sauteed B	4.375	107	4	10	329	0	0	0	0	0	0	0	0	1	0	1	0	0	0
27	Grouper w	4.375	336	44	16	413	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	Better-Tha	2.5	145	3	6	208	0	0	0	0	0	0	0	0	0	0	1	0	0	0

The translate data to visualization aspects:

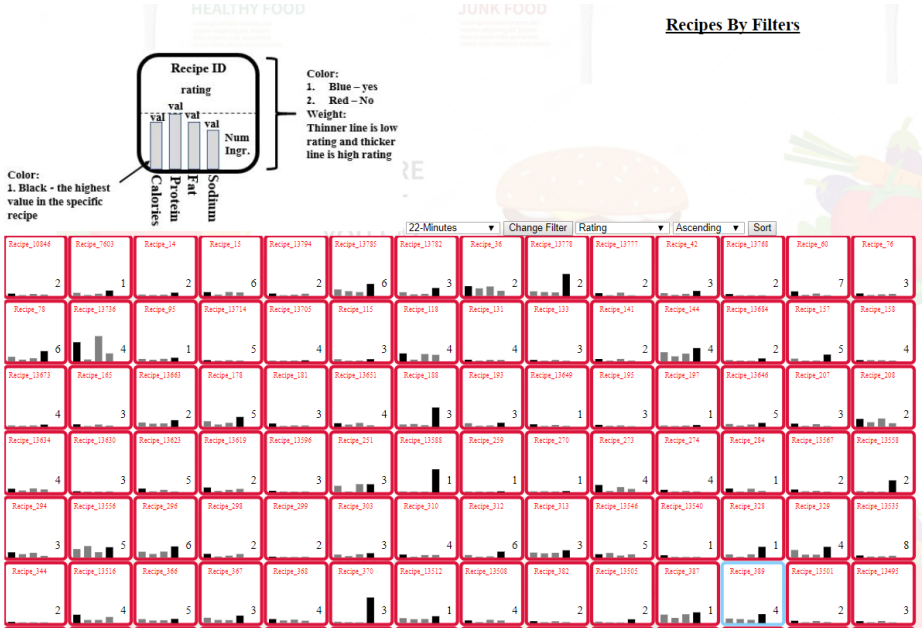
- Every recipe – a rectangle **(shape)**
- Border weight – the rating of the recipe **(weight)**
- Recipe ID – **(label)**
- Rating – **(label)**
- Calories, protein, fat, sodium – bar plot **(shape + position + label)**
- Ingredients – **(label)**
- Filter – **(color)**
- Higher value between four categories in the recipe – black bar **(color)**
- The heights of the bars – normalized related to the all other recipes **(y-position)**



Interaction:

- Hover on the rectangle – shows the rating value and the border color is yellow
- Hover on the bars – shows the value of the specific bar
- Clicking on the rectangle – alert with all the relevant metadata of the recipe

The visualization:



Pros and Cons:

- Time:

  + Filters are very fast

  - Takes time for creating the whole matrix

- Insights:

  + Can find the best recipe under some requirements

  + Feel the data by filters

- Essence:

+ Seeing the big picture by sorting

+ All the bars are related to the all other recipes

- Must using the interaction for scrolling down and up for finding the wanted recipe

- Confidence:

  + Uses all the relevant data and show it all in the same time

General speaking:

To sum up the pros and cons of this visualization to my project, the visualization allows the user to explore the data and define the specific recipe he wants under some requirements. All the data of the recipe is in little square consists all the data the user needs. The big drawback of it is that it covers very big area and it hard to scroll down to see everything. The sort and the filter supposed to help find the pattern at the beginning of the page.

Interesting Insights:

- As expected, there are only 24 recipes for Passover, but the average rating is not as expected (3.54 and not under 1)
- Most of the desserts have high rating
- All the recipes that consists above 2589 calories are not Healthy
- Most of the breakfast are Kosher but consists high cholesterol and high levels of sugar.