DietDupe

Explorative data analysis

Piotr Kaszubski Antoni Solarski Nina Żukowska

Poznań University of Technology

Monday, November 13, 2023

Choosing the dataset

1. FlavorGraph [github]
   - `FlavorGraph_node_embedding.pickle`
   - `nodes_191120.csv`
2. The Nutritional Content of Food [kaggle]
   - `ABBREV.csv`
3. (future work?) Vegan/non-vegan dataset

Dataset – FlavorGraph

```python
for k, v in data.items():
    print(k, v)
    break
```

```
81 [ 0.06278703  0.03142117  0.01589981  0.15482351 -0.20670895 -0.02320386
  0.18804736 -0.03847743  0.08789175  0.14633657 -0.03981635  0.0688429
 -0.14873533 -0.07678423 -0.02092623  0.25386092 -0.0954157  -0.01681494
 -0.11518956  0.18274969  0.09485493 -0.41762635  0.20165439 -0.12645648
 -0.01135357  0.07055754  0.3084545   0.02781115  0.04194851 -0.11863893
 -0.19511323 -0.14121613 -0.1497203  -0.05314291 -0.3111925  -0.13476767
 -0.15796486  0.2069276  -0.08805668 -0.30796188  0.09973499  0.09277296
 -0.16415025 -0.11142119  0.12177276  0.15759766 -0.17777048 -0.02054236
  0.03809059  0.2022568   0.17120546 -0.13465042 -0.23092705  0.12159029
 -0.19453536 -0.00532992 -0.1522021   0.01335487 -0.13338162 -0.2872366
  0.02461791 -0.10121837  0.04762103 -0.05549576 -0.1204869   0.09709413
 -0.05324109 -0.24248974 -0.1294401   0.01897155 -0.02538477  0.01806459
  0.03920558  0.32651514  0.2072027  -0.17189099 -0.26355118  0.06044648
  0.09988082 -0.00153702  0.04054973 -0.24209592  0.02980894  0.22471131
 -0.19577554 -0.05329362  0.10919486  0.0293193   0.24686033 -0.09472723
 0.1229382   0.12351561 -0.16917424  0.05163106  0.00158128  0.17401817
```

Figure: `FlavorGraph_node_embedding.pickle`

Dataset – FlavorGraph

```
nodes.head()
```

	node_id	name	id	node_type	is_hub
0	0	1%_fat_buttermilk	NaN	ingredient	no_hub
1	1	1%_fat_cottage_cheese	NaN	ingredient	no_hub
2	3	10%_cream	NaN	ingredient	no_hub
3	4	100%_bran	NaN	ingredient	no_hub
4	5	10_inch_flour_tortilla	NaN	ingredient	no_hub

Figure: `nodes_191120.csv`

Dataset – The Nutritional Content of Food

	Column name	Description
0	index	The index of the row
1	NDB_No	The National Diet and Nutrition Survey number of the food
2	Shrt_Desc	The short description of the food
3	Energ_Kcal	The amount of energy in the food
4	Lut+Zea_ (µg)	Lutein and Zeaxanthin (µg)
5	Vit_E_(mg)	Vitamin E (mg)
6	Vit_D_µg	Vitamin D (µg)
7	Vit_D_IU	Vitamin D (IU)
8	Vit_K_(µg)	Vitamin K (µg)
9	FA_Sat_(g)	Saturated Fat (g)
10	FA_Mono_(g)	Monounsaturated Fat (g)
11	FA_Poly_(g)	Polyunsaturated Fat (g)
12	Cholestrl_(mg)	Cholesterol (mg)
13	GmWt_1	Gram Weight 1

Figure: `ABBREV.csv` columns (brief)

Dataset – The Nutritional Content of Food

```
nutri_data.head()
```

	index	NDB_No	Shrt_Desc	Water_(g)	Energ_Kcal	Protein_(g)	Lipid_Tot_(g)	Ash_(g)	Carbohydrt_(g)	Fiber_TD_(g)	...	Vit_K_(µg)	FA_Sat_(g)	FA_Mono_(g)
0	0	1001	BUTTER,WITH SALT	15.87	717	0.85	81.11	2.11	0.06	0.0	...	7.0	51.368	21.021
1	1	1002	BUTTER,WHIPPED,W/ SALT	16.72	718	0.49	78.30	1.62	2.87	0.0	...	4.6	45.390	19.874
2	2	1003	BUTTER OIL,ANHYDROUS	0.24	876	0.28	99.48	0.00	0.00	0.0	...	8.6	61.924	28.732
3	3	1004	CHEESE,BLUE	42.41	353	21.40	28.74	5.11	2.34	0.0	...	2.4	18.669	7.778
4	4	1005	CHEESE,BRICK	41.11	371	23.24	29.68	3.18	2.79	0.0	...	2.5	18.764	8.598

5 rows × 54 columns

Figure: `ABBREV.csv`

Matching the two datasets

Method 1: BERT

1. Use `BertTokenizer` to transform both datasets into same-space embeddings.
2. Match terms based on cosine similarity.

```
nodes.head(20)
```

	node_id	name	id	node_type	is_hub	embeddings	best_match	similarity_of_best_match
0	0.0	1%_fat_buttermilk	NaN	ingredient	no_hub	[-0.10600116, 0.047149494, 0.10841199, 0.07235...	MILK,BUTTERMILK,DRIED	0.859678
1	1.0	1%_fat_cottage_cheese	NaN	ingredient	no_hub	[-0.015829312, 0.09736368, -0.0006226096, 0.13...	CHEESE,COTTAGE,LOWFAT,1% MILKFAT	0.84297
2	3.0	10%_cream	NaN	ingredient	no_hub	[-0.10132008, 0.033723958, 0.064727835, 0.1566...	TURTLE,GREEN,RAW	0.738032
3	4.0	100%_bran	NaN	ingredient	no_hub	[-0.10309663, 0.03204953, 0.08858223, 0.105722...	TURTLE,GREEN,RAW	0.745624
4	5.0	10_inch_flour_tortilla	NaN	ingredient	no_hub	[-0.09346332, 0.120890595, 0.10606088, 0.10007...	Tortilla chips, yellow, plain, salted	0.864151
5	7.0	12_inch_pizza_crust	NaN	ingredient	no_hub	[0.019675368, 0.0335593, -0.05435514, 0.112564...	PIZZA HUT 12" PEPPERONI PIZZA,PAN CRUST	0.822786
6	9.0	18%_table_cream	NaN	ingredient	no_hub	[-0.17582406, 0.14053129, -0.022183008, 0.0534...	PORK,GROUND,84% LN / 16% FAT,RAW	0.781034
7	10.0	2%_buttermilk	NaN	ingredient	no_hub	[-0.07227368, 0.033502933, 0.037231416, 0.1503...	MILK,BUTTERMILK,DRIED	0.871265
8	11.0	2%_cheddar_cheese	NaN	ingredient	no_hub	[0.12324965, -0.26437616, -0.09264475, 0.06749...	CHEESE,CHEDDAR	0.879752

Matching the two datasets

Method 2: `food2vec`

1. Use `food2vec.semantic_nutrition.Estimator` to transform both datasets into same-space embeddings.
2. Match terms based on cosine similarity.

```
nodes_f2v.head(20)
```

	node_id	name	id	node_type	is_hub	embeddings	best_match	similarity_of_best_match
0	0.0	1%_fat_buttermilk	NaN	ingredient	no_hub	[-0.10600116, 0.047149494, 0.10841199, 0.07235...	MARGARINE-LIKE SPRD W/ YOGURT,70% FAT,STK,W/ SALT	0.792741
1	1.0	1%_fat_cottage_cheese	NaN	ingredient	no_hub	[-0.015829312, 0.09736368, -0.0006226096, 0.13...	CHEESE,COTTAGE,CRMD,W/FRUIT	0.871232
2	3.0	10%_cream	NaN	ingredient	no_hub	[-0.10132008, 0.033723958, 0.064727835, 0.1566...	CHEESE,CREAM	0.812511
3	4.0	100%_bran	NaN	ingredient	no_hub	[-0.10309663, 0.03204953, 0.08858223, 0.105722...	CORN BRAN,CRUDE	0.859974
4	5.0	10_inch_flour_tortilla	NaN	ingredient	no_hub	[-0.09346332, 0.120890595, 0.10606088, 0.10007...	WHEAT FLR,WHITE,TORTILLA MIX,ENR	0.76415
5	7.0	12_inch_pizza_crust	NaN	ingredient	no_hub	[0.019675368, 0.0335593, -0.05435514, 0.112564...	PAPA JOHN'S 14" CHS PIZZA,ORIGINAL CRUST	0.872573
6	9.0	18%_table_cream	NaN	ingredient	no_hub	[-0.17582406, 0.14053129, -0.022183008, 0.0534...	CREAM,FLUID,LT (COFFEE CRM OR TABLE CRM)	0.937748
7	10.0	2%_buttermilk	NaN	ingredient	no_hub	[-0.07227368, 0.033502933, 0.037231416, 0.1503...	MILK,BUTTERMILK,DRIED	0.833767
8	11.0	2%_cheddar_cheese	NaN	ingredient	no_hub	[0.12324965, -0.26437616, -0.09264475, 0.06749...	CHEESE,CHEDDAR	1.0

Matching the two datasets

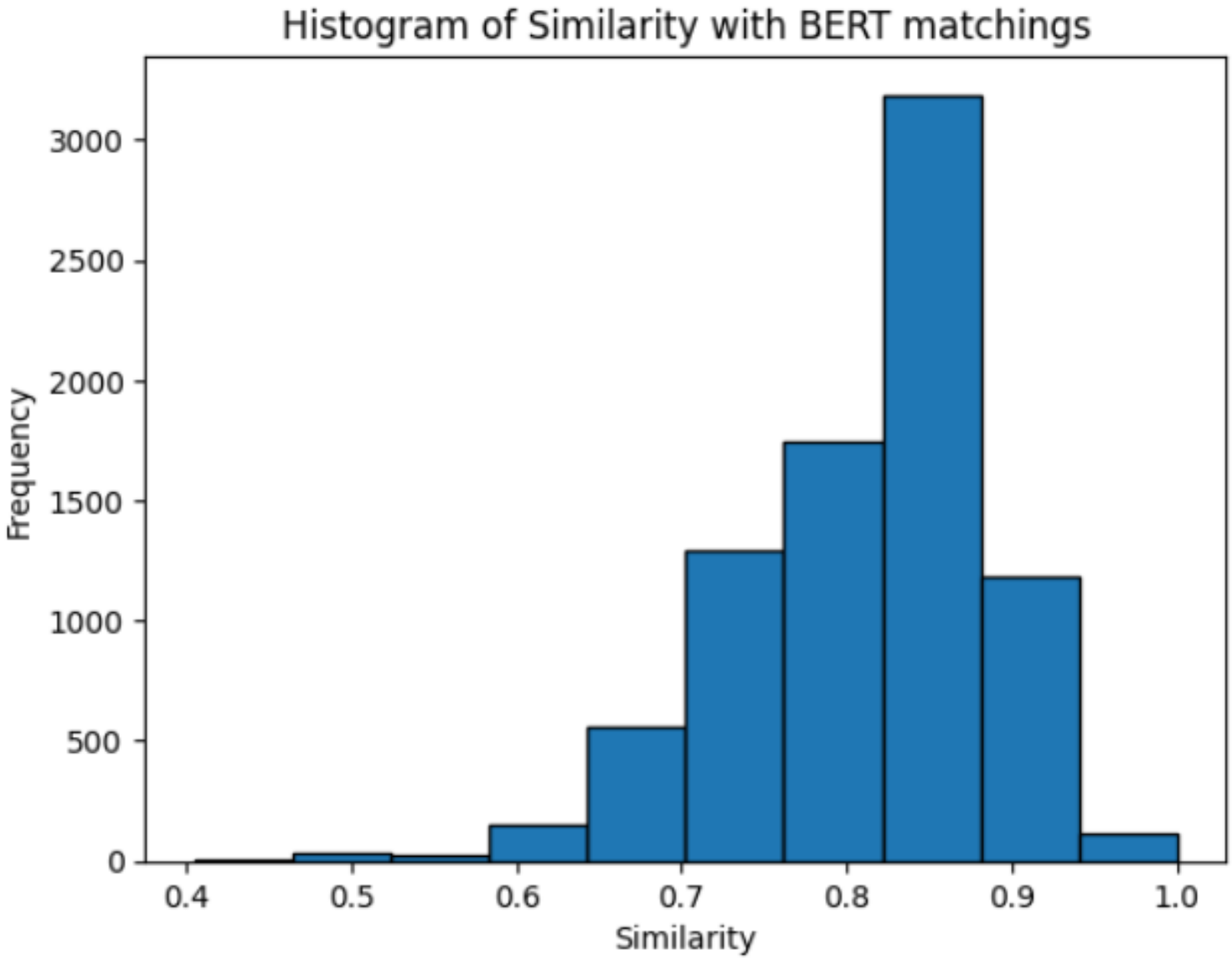
Combining both methods

1. Fill nodes from `food2vec` with nodes from `BERT`.

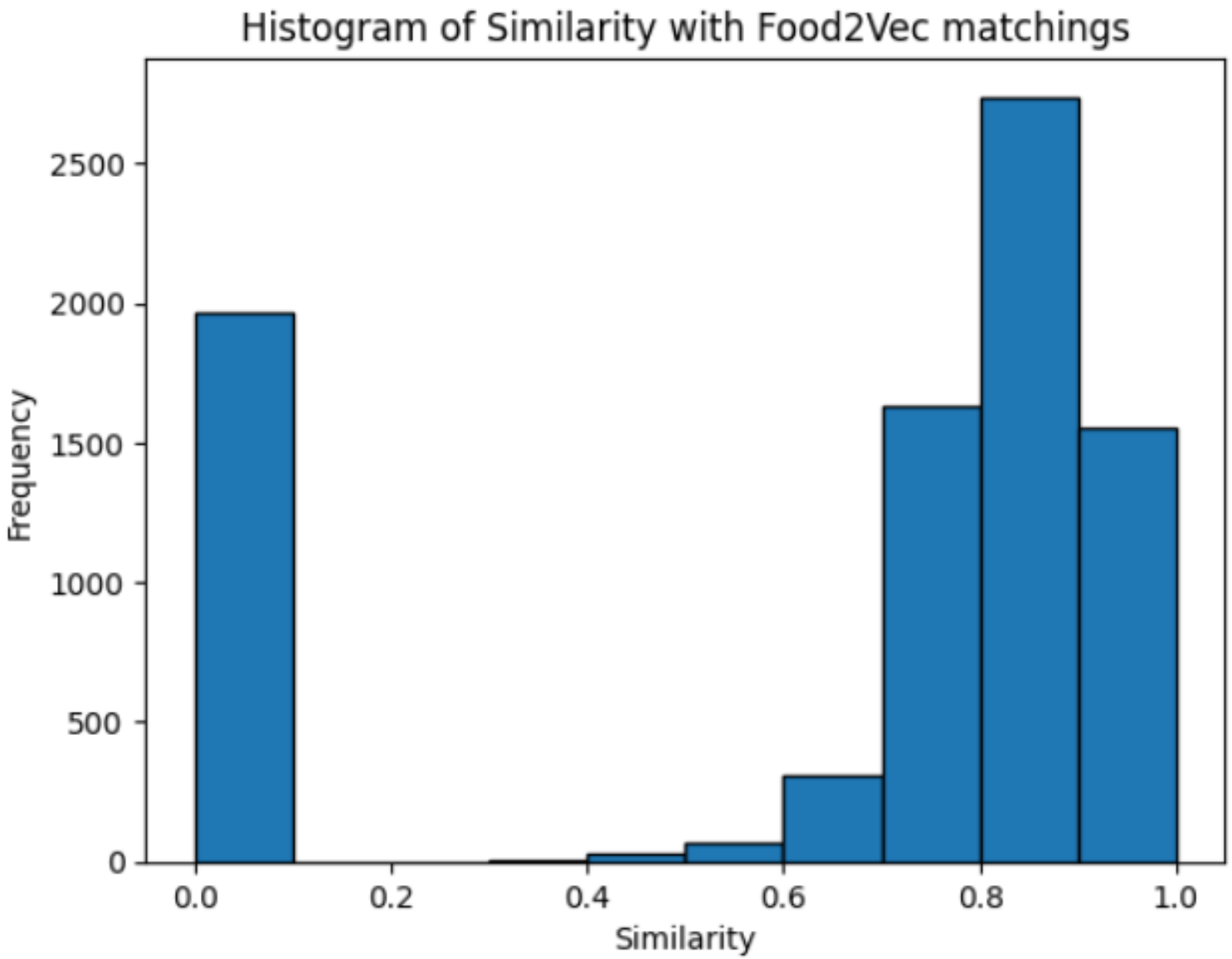
```
nodes_f2v[nodes_f2v['similarity_of_best_match'] == 0.0].head(30)
```

	node_id	name	id	node_type	is_hub	embeddings	best_match	similarity_of_best_match	bert_embeddings_similarity
33	38.0	abalone	NaN	ingredient	hub	[0.21432944, -0.24550392, -0.14755642, 0.24025...	CRAB,BLUE,CANNED	0.0	0.804253
34	39.0	absinthe	NaN	ingredient	no_hub	[-0.1563259, -0.12958135, -0.105907924, 0.1457...	SPICES,CARDAMOM	0.0	0.734083
36	44.0	achiote	NaN	ingredient	no_hub	[-0.20564497, 0.0021581268, -0.017484589, 0.15...	NATTO	0.0	0.759630
44	53.0	active_starter	NaN	ingredient	no_hub	[-0.18706283, 0.15871908, 0.01949941, 0.125311...	CHEWING GUM	0.0	0.670433
47	56.0	advocaat	NaN	ingredient	no_hub	[0.06339699, -0.13852063, -0.029949985, 0.0742...	CELTUCE,RAW	0.0	0.667664
49	58.0	agar	NaN	ingredient	hub	[0.091211654, -0.20150454, -0.26364943, 0.0497...	HONEY	0.0	0.760247
50	59.0	agar_agar	NaN	ingredient	no_hub	[-0.02377455, -0.032742925, -0.016038483, 0.01...	SEAWEED,AGAR,RAW	0.0	0.704856
60	69.0	ajinomoto	NaN	ingredient	no_hub	[-0.10433134, 0.021288093, 0.22506452, 0.17667...	OKARA	0.0	0.696734
61	70.0	ajwain	NaN	ingredient	no_hub	[-0.19063282, -0.05176021, 0.048555177, 0.1778...	DATES,DEGLET NOOR	0.0	0.766456
64	73.0	alcaparrado	NaN	ingredient	no_hub	[-0.1326251, 0.040735394, -0.09200989, -0.0749...	DULCE DE LECHE	0.0	0.782369
77	86.0	allspice	NaN	ingredient	hub	[0.12725407, -0.15135208, -0.30460206, 0.15687...	ALLSPICE,GROUND	0.0	0.851627

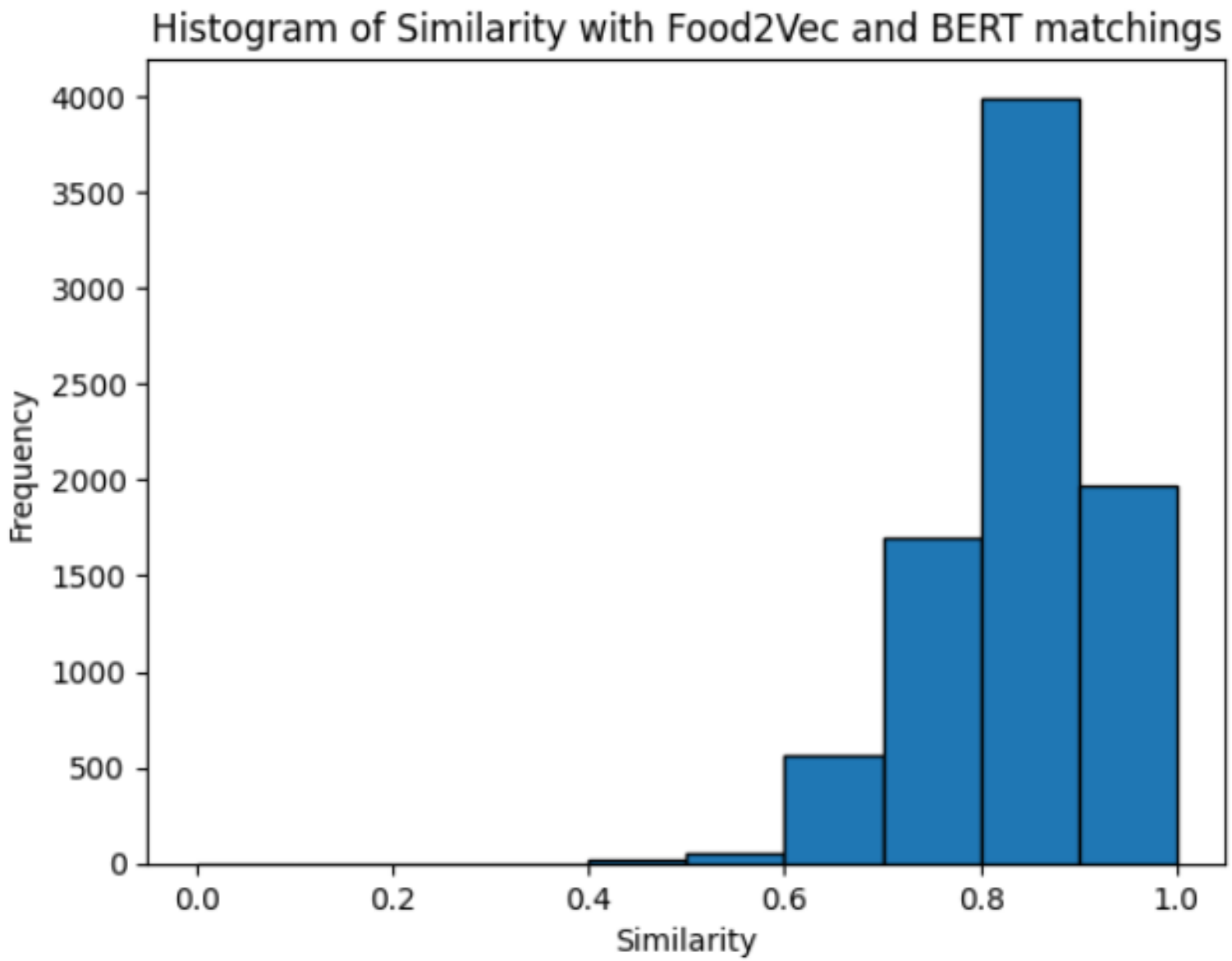
Data exploration



Data exploration



Data exploration



Data exploration

	average	standard_deviation
Calcium_(mg)	76.738214	203.527453
Iron_(mg)	2.699674	5.687560
Magnesium_(mg)	35.295988	57.416785
Phosphorus_(mg)	165.142126	204.704214
Potassium_(mg)	279.472740	375.483729
Sodium_(mg)	312.495923	943.431341
Zinc_(mg)	2.117438	3.437209
Copper_mg)	0.195984	0.582596
Manganese_(mg)	0.658156	7.248609
Vit_C_(mg)	9.231134	68.854696
Thiamin_(mg)	0.223134	0.523752
Riboflavin_(mg)	0.252237	0.449461
Niacin_(mg)	3.657721	4.823819
Panto_Acid_mg)	0.650989	1.413303
Vit_B6_(mg)	0.291531	0.485057
Choline_Tot_ (mg)	43.596230	63.109904
Vit_E_(mg)	1.331518	4.640706
Cholestrl_(mg)	40.613246	119.869371

Figure: Summary of micronutrient columns