CSE564 Visualization

Yinlong Su, #110461173

April 10, 2016

Raw Data Report

1 CSV Report

There are 65,503 rows and 159 columns in raw data.

2 Column Report

Many columns are incomplete. Some missing data is unrecoverable, while some can be filled as default value. For example, column "lignoceric_acid_100g" are empty so intuitively we know that the food has no lignoceric acid.

Here we roughly divide the data type into 5 types: Numeric, Text, Time, URL and Unknown. See table 1.

Table 1: Column Report Table

Column Name	Type	Empty/Invalid
code	numericType	17 (0.03%)
url	urlType	17 (0.03%)
creator	textType	68 (0.1%)
created_t	timeType	3 (0.0%)
created_datetime	timeType	7 (0.01%)
last_modified_t	timeType	0 (0.0%)
last_modified_datetime	timeType	0 (0.0%)
product_name	textType	5230 (7.98%)
generic_name	textType	31290 (47.77%)
quantity	textType	9218 (14.07%)
packaging	textType	15425 (23.55%)
packaging_tags	textType	15425 (23.55%)
brands	textType	9303 (14.2%)
brands_tags	textType	9304 (14.2%)
categories	textType	15608 (23.83%)
categories_tags	textType	15623 (23.85%)
categories_en	textType	15607 (23.83%)
origins	textType	50138 (76.54%)
origins_tags	textType	50175 (76.6%)
manufacturing_places	textType	44435 (67.84%)
manufacturing_places_tags	textType	44441 (67.85%)
labels	textType	40756 (62.22%)

Continued on next page

Table 1 – continued from previous page

Column Name	Type	Empty/Invalid
labels_tags	textType	40709 (62.15%)
labels_en	textType	40693 (62.12%)
emb_codes	textType	46138 (70.44%)
emb_codes_tags	textType	46141 (70.44%)
first_packaging_code_geo	textType	53278 (81.34%)
cities	textType	65487 (99.98%)
cities_tags	textType	52241 (79.75%)
purchase_places	textType	28271 (43.16%)
stores	textType	33419 (51.02%)
countries	textType	211 (0.32%)
countries_tags	textType	211 (0.32%)
countries_en	textType	211 (0.32%)
ingredients_text	textType	21822 (33.31%)
allergens	textType	51966 (79.33%)
allergens_en	textType	65486 (99.97%)
traces	textType	50631 (77.3%)
traces_tags	textType	50649 (77.32%)
traces_en	textType	50633 (77.3%)
serving_size	textType	44329 (67.67%)
no_nutriments	unknownType	65503 (100.0%)
additives_n	numericType	21839 (33.34%)
additives	textType	21857 (33.37%)
additives_tags	textType	41838 (63.87%)
additives_en	textType	41838 (63.87%)
ingredients_from_palm_oil_n	numericType	21839 (33.34%)
ingredients_from_palm_oil	textType	65503 (100.0%)
ingredients_from_palm_oil_tags	textType	63122 (96.37%)
ingredients_that_may_be_from_palm_oil_n	numericType	21839 (33.34%)
ingredients_that_may_be_from_palm_oil	textType	65503 (100.0%)
ingredients_that_may_be_from_palm_oil_tags	textType	60981 (93.1%)
nutrition_grade_uk	textType	65503 (100.0%)
nutrition_grade_fr	textType	34209 (52.23%)
pnns_groups_1	textType	12997 (19.84%)
pnns_groups_2	textType	11087 (16.93%)
states	textType	88 (0.13%)
states_tags	textType	88 (0.13%)
states_en	textType	88 (0.13%)
main_category	textType	15640 (23.88%)
main_category_en	textType	15640 (23.88%)
image_url	urlType	4304 (6.57%)
image_small_url	urlType	4304 (6.57%)
energy_100g	numericType	29129 (44.47%)
energy_from_fat_100g	numericType	64770 (98.88%)
fat_100g	numericType	29141 (44.49%)
saturated_fat_100g	numericType	33074 (50.49%)
butyric_acid_100g	numericType	65503 (100.0%)
caproic_acid_100g	numericType	65503 (100.0%)
caprylic_acid_100g	numericType	65502 (100.0%)

Continued on next page

Table 1 – continued from previous page

Column Name	Type	Empty/Invalid
capric_acid_100g	numericType	65502 (100.0%)
lauric_acid_100g	numericType	65500 (100.0%)
myristic_acid_100g	numericType	65502 (100.0%)
palmitic_acid_100g	numericType	65502 (100.0%)
stearic_acid_100g	numericType	65502 (100.0%)
arachidic_acid_100g	numericType	65486 (99.97%)
behenic_acid_100g	numericType	65487 (99.98%)
lignoceric_acid_100g	numericType	65503 (100.0%)
cerotic_acid_100g	numericType	65503 (100.0%)
montanic_acid_100g	numericType	65503 (100.0%)
melissic_acid_100g	numericType	65503 (100.0%)
monounsaturated_fat_100g	numericType	63946 (97.62%)
polyunsaturated_fat_100g	numericType	63933 (97.6%)
omega_3_fat_100g	numericType	64976 (99.2%)
alpha_linolenic_acid_100g	numericType	65374 (99.8%)
eicosapentaenoic_acid_100g	numericType	65470 (99.95%)
docosahexaenoic_acid_100g	numericType	65446 (99.91%)
omega_6_fat_100g	numericType	65366 (99.79%)
linoleic_acid_100g	numericType	65406 (99.85%)
arachidonic_acid_100g	numericType	65498 (99.99%)
gamma_linolenic_acid_100g	numericType	65486 (99.97%)
dihomo_gamma_linolenic_acid_100g	numericType	65487 (99.98%)
omega_9_fat_100g	numericType	65487 (99.98%)
oleic_acid_100g	numericType	65496 (99.99%)
elaidic_acid_100g	numericType	65503 (100.0%)
gondoic_acid_100g	numericType	65495 (99.99%)
mead_acid_100g	numericType	65503 (100.0%)
erucic_acid_100g	numericType	65503 (100.0%)
nervonic_acid_100g	numericType	65503 (100.0%)
trans_fat_100g	numericType	64275 (98.13%)
cholesterol_100g	numericType	64112 (97.88%)
carbohydrates_100g	numericType	29438 (44.94%)
sugars_100g	numericType	32864 (50.17%)
sucrose_100g	numericType	65495 (99.99%)
glucose_100g	numericType	65498 (99.99%)
fructose_100g	numericType	65483 (99.97%)
lactose_100g	numericType	65362 (99.78%)
maltose_100g	numericType	65501 (100.0%)
maltodextrins_100g	numericType	65496 (99.99%)
starch_100g	numericType	65287 (99.67%)
polyols_100g	numericType	65253 (99.62%)
fiber_100g	numericType	42957 (65.58%)
proteins_100g	numericType	29573 (45.15%)
casein_100g	numericType	65488 (99.98%)
serum_proteins_100g	numericType	65495 (99.99%)
nucleotides_100g	numericType	65500 (100.0%)
salt_100g	numericType	32595 (49.76%)
sodium_100g	numericType	32605 (49.78%)

Continued on next page

Table 1 – continued from previous page

Column Name	Type	Empty/Invalid
alcohol_100g	numericType	63084 (96.31%)
vitamin_a_100g	numericType	64142 (97.92%)
beta_carotene_100g	numericType	65494 (99.99%)
vitamin_d_100g	numericType	64921 (99.11%)
vitamin_e_100g	numericType	64746 (98.84%)
vitamin_k_100g	numericType	65444 (99.91%)
vitamin_c_100g	numericType	63595 (97.09%)
vitamin_b1_100g	numericType	64632 (98.67%)
vitamin_b2_100g	numericType	64782 (98.9%)
vitamin_pp_100g	numericType	64772 (98.88%)
vitamin_b6_100g	numericType	64800 (98.93%)
vitamin_b9_100g	numericType	64766 (98.87%)
vitamin_b12_100g	numericType	64912 (99.1%)
biotin_100g	numericType	65310 (99.71%)
pantothenic_acid_100g	numericType	65096 (99.38%)
silica_100g	numericType	65475 (99.96%)
bicarbonate_100g	numericType	65441 (99.91%)
potassium_100g	numericType	65006 (99.24%)
chloride_100g	numericType	65397 (99.84%)
calcium_100g	numericType	62554 (95.5%)
phosphorus_100g	numericType	64913 (99.1%)
iron_100g	numericType	63650 (97.17%)
magnesium_100g	numericType	64706 (98.78%)
zinc_100g	numericType	65258 (99.63%)
copper_100g	numericType	65407 (99.85%)
manganese_100g	numericType	65418 (99.87%)
fluoride_100g	numericType	65448 (99.92%)
selenium_100g	numericType	65429 (99.89%)
chromium_100g	numericType	65488 (99.98%)
molybdenum_100g	numericType	65498 (99.99%)
iodine_100g	numericType	65388 (99.82%)
caffeine_100g	numericType	65467 (99.95%)
taurine_100g	numericType	65487 (99.98%)
ph_100g	numericType	65468 (99.95%)
fruits_vegetables_nuts_100g	numericType	64474 (98.43%)
collagen_meat_protein_ratio_100g	numericType	65394 (99.83%)
cocoa_100g	numericType	65071 (99.34%)
chlorophyl_100g	numericType	65503 (100.0%)
carbon_footprint_100g	numericType	65323 (99.73%)
nutrition_score_fr_100g	numericType	34209 (52.23%)
nutrition_score_uk_100g	numericType	34209 (52.23%)

Figure 1 shows the level of integrities of the columns.

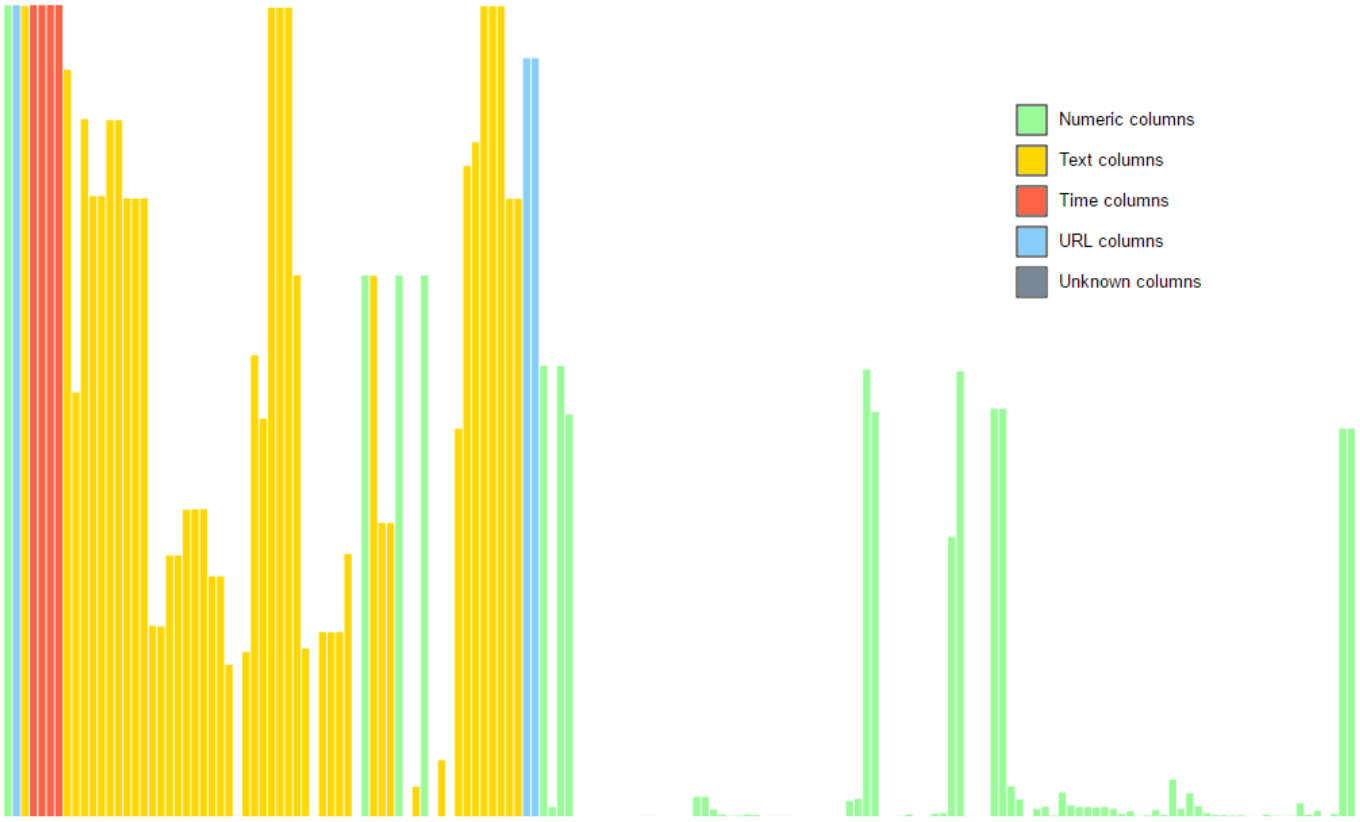

Figure 1: Interities of Raw Columns

3 Numeric Columns Report

Table 2 shows the report of numeric columns. I ommited the average values, see "raw.analysis.txt" if you need them.

Table 2: Numeric Column Report Table

Column Name	Min	Max	Empty/Invalid
additives_n	0.0	26.0	21839 (33.34%)
ingredients_from_palm_oil_n	0.0	2.0	21839 (33.34%)
ingredients_that_may_be_from_palm_oil_n	0.0	6.0	21839 (33.34%)
energy_100g	0.0	4134.0	29129 (44.47%)
energy_from_fat_100g	0.0	3740.0	64770 (98.88%)
fat_100g	0.0	101.0	29141 (44.49%)
saturated_fat_100g	0.0	100.0	33074 (50.49%)
butyric_acid_100g	nan	nan	65503 (100.0%)
caproic_acid_100g	nan	nan	65503 (100.0%)
caprylic_acid_100g	7.4	7.4	65502 (100.0%)
capric_acid_100g	6.2	6.2	65502 (100.0%)
lauric_acid_100g	0.04473	49.3	65500 (100.0%)
myristic_acid_100g	18.9	18.9	65502 (100.0%)
palmitic_acid_100g	8.1	8.1	65502 (100.0%)
stearic_acid_100g	3.0	3.0	65502 (100.0%)
arachidic_acid_100g	0.064	15.4	65486 (99.97%)

Continued on next page

Table 2 – continued from previous page

Column Name	Min	Max	Empty/Invalid
behenic_acid_100g	5.5	14.6	65487 (99.98%)
lignoceric_acid_100g	nan	nan	65503 (100.0%)
cerotic_acid_100g	nan	nan	65503 (100.0%)
montanic_acid_100g	nan	nan	65503 (100.0%)
melissic_acid_100g	nan	nan	65503 (100.0%)
monounsaturated_fat_100g	0.0	80.0	63946 (97.62%)
polyunsaturated_fat_100g	0.0	74.0	63933 (97.6%)
omega_3_fat_100g	0.00015	38.2	64976 (99.2%)
alpha_linolenic_acid_100g	0.0	47.0	65374 (99.8%)
eicosapentaenoic_acid_100g	0.048	85.0	65470 (99.95%)
docosahexaenoic_acid_100g	0.044	12.0	65446 (99.91%)
omega_6_fat_100g	0.05	71.0	65366 (99.79%)
linoleic_acid_100g	0.09	25.0	65406 (99.85%)
arachidonic_acid_100g	0.007	0.09	65498 (99.99%)
gamma_linolenic_acid_100g	0.095	0.2032	65486 (99.97%)
dihomo_gamma_linolenic_acid_100g	0.05	0.08	65487 (99.98%)
omega_9_fat_100g	1.0	75.0	65487 (99.98%)
oleic_acid_100g	1.08	70.0	65496 (99.99%)
elaidic_acid_100g	nan	nan	65503 (100.0%)
gondoic_acid_100g	1e-06	1.25e-06	65495 (99.99%)
mead_acid_100g	nan	nan	65503 (100.0%)
erucic_acid_100g	nan	nan	65503 (100.0%)
nervonic_acid_100g	nan	nan	65503 (100.0%)
trans_fat_100g	0.0	30.3	64275 (98.13%)
cholesterol_100g	0.0	0.432	64112 (97.88%)
carbohydrates_100g	0.0	139.0	29438 (44.94%)
sugars_100g	-0.5	105.0	32864 (50.17%)
sucrose_100g	0.0	8.4	65495 (99.99%)
glucose_100g	0.1	2.0	65498 (99.99%)
fructose_100g	0.1	101.0	65483 (99.97%)
lactose_100g	0.0	58.5	65362 (99.78%)
maltose_100g	0.1	22.0	65501 (100.0%)
maltodextrins_100g	1.8	27.5	65496 (99.99%)
starch_100g	0.0	87.8	65287 (99.67%)
polyols_100g	0.0	100.0	65253 (99.62%)
fiber_100g	0.0	94.8	42957 (65.58%)
proteins_100g	0.0	86.0	29573 (45.15%)
casein_100g	0.92	10.2	65488 (99.98%)
serum_proteins_100g	0.3	5.8	65495 (99.99%)
nucleotides_100g	0.018	0.024	65500 (100.0%)
salt_100g	0.0	254.0	32595 (49.76%)
sodium_100g	0.0	100.0	32605 (49.78%)
alcohol_100g	0.0	97.9	63084 (96.31%)
vitamin_a_100g	0.0	26.7	64142 (97.92%)
beta_carotene_100g	0.000812	0.26	65494 (99.99%)
vitamin_d_100g	0.0	7.5	64921 (99.11%)
vitamin_e_100g	0.0	15.1	64746 (98.84%)
vitamin_k_100g	0.0	0.0334	65444 (99.91%)

Continued on next page

Table 2 – continued from previous page

Column Name	Min	Max	Empty/Invalid
vitamin_c_100g	0.0	100.0	63595 (97.09%)
vitamin_b1_100g	1e-05	0.169	64632 (98.67%)
vitamin_b2_100g	0.0	30.0	64782 (98.9%)
vitamin_pp_100g	2.4e-06	15.9	64772 (98.88%)
vitamin_b6_100g	0.0	0.215	64800 (98.93%)
vitamin_b9_100g	0.0	4.0	64766 (98.87%)
vitamin_b12_100g	0.0	50.0	64912 (99.1%)
biotin_100g	-2.0	5.0	65310 (99.71%)
pantothenic_acid_100g	-2.0	5.0	65096 (99.38%)
silica_100g	8.2e-06	0.0362	65475 (99.96%)
bicarbonate_100g	6.3e-06	1.1	65441 (99.91%)
potassium_100g	0.0	34.6	65006 (99.24%)
chloride_100g	2e-06	0.5	65397 (99.84%)
calcium_100g	0.0	69.5	62554 (95.5%)
phosphorus_100g	0.0012	84.9	64913 (99.1%)
iron_100g	0.0	19.2	63650 (97.17%)
magnesium_100g	5e-07	11.5	64706 (98.78%)
zinc_100g	0.0001	4.0	65258 (99.63%)
copper_100g	1.67e-05	0.35	65407 (99.85%)
manganese_100g	3e-06	0.7	65418 (99.87%)
fluoride_100g	0.0	0.56	65448 (99.92%)
selenium_100g	7e-07	0.03	65429 (99.89%)
chromium_100g	7.06e-06	0.0001	65488 (99.98%)
molybdenum_100g	7e-06	4.5e-05	65498 (99.99%)
iodine_100g	3.6e-09	0.0147	65388 (99.82%)
caffeine_100g	0.003	32.0	65467 (99.95%)
taurine_100g	0.0018	0.423	65487 (99.98%)
ph_100g	0.005	7.9	65468 (99.95%)
fruits_vegetables_nuts_100g	1.4	100.0	64474 (98.43%)
collagen_meat_protein_ratio_100g	8.0	25.0	65394 (99.83%)
cocoa_100g	6.3	100.0	65071 (99.34%)
chlorophyl_100g	nan	nan	65503 (100.0%)
carbon_footprint_100g	0.0	2842.0	65323 (99.73%)
nutrition_score_fr_100g	-14.0	35.0	34209 (52.23%)
nutrition_score_uk_100g	-14.0	33.0	34209 (52.23%)

4 Text Group Columns Report

Table 3 shows the number of different values on selected text columns. Some of the columns are actually multiset of text information. The number is too large to be reasonably group. Further processing is needed.

Table 3: Text Group Column Report Table

Column Name	Group #	Empty/Invalid
creator	1905	68 (0.1%)
Continued on next page		

Table 3 – continued from previous page

Column Name	Group #	Empty/Invalid
brands	17172	9303 (14.2%)
brands_tags	14097	9304 (14.2%)
categories	23878	15608 (23.83%)
categories_tags	14063	15623 (23.85%)
origins	3655	50138 (76.54%)
origins_tags	3313	50175 (76.6%)
manufacturing_places	4424	44435 (67.84%)
manufacturing_places_tags	4033	44441 (67.85%)
countries	722	211 (0.32%)
countries_tags	373	211 (0.32%)
countries_en	372	211 (0.32%)
nutrition_grade_uk	0	65503 (100.0%)
nutrition_grade_fr	5	34209 (52.23%)
main_category_en	2665	15640 (23.88%)

5 Contact Information

If there is any problem please contact me.

Name: Yinlong Su
SBU ID: 110461173
Email: yinlsu@cs.stonybrook.edu