Project 2: Food Nutrition and Open-Source Databases

By Yubo Zhang, Carlee Price, and Nikki Haas

August 2016

Python W-18

Introduction

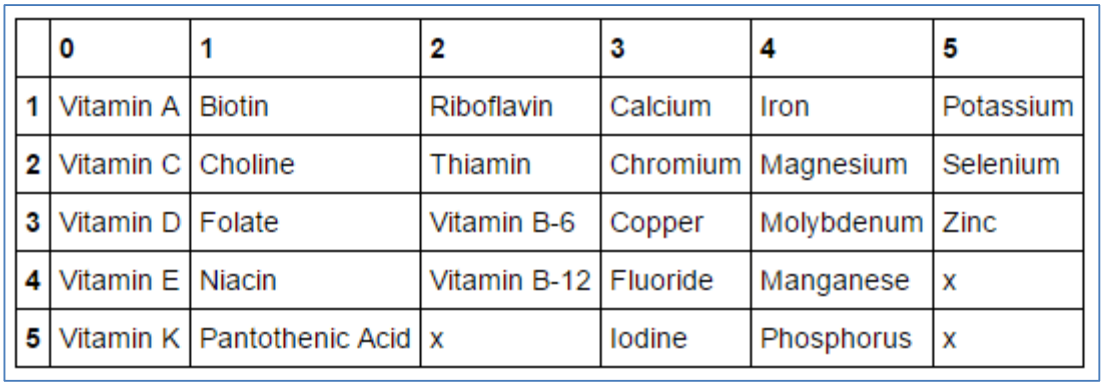
We used Open Food Facts dataset for our final project. The database is a "free, open and collaborative database of food products from the entire world". The database includes 150 fields (both numerical and descriptive) and over 90,000 items. For this project, we will work with the subset of data which includes products available in the US (2,800 items). We'll ask of this data questions related to nutritional content, and for that purpose will pull in data from additional sources including the IOM (Institute of Medicine) and the FDA (Food and Drug Administration).

The data was groomed and manipulated using the Pandas, Numpy, Matplotlib, Regular Expressions and SciPy modules available for Python.

Question 1: Is commercially available foods nutritious enough to meet both the daily recommended values for vitamins and minerals while staying in the daily caloric budget?

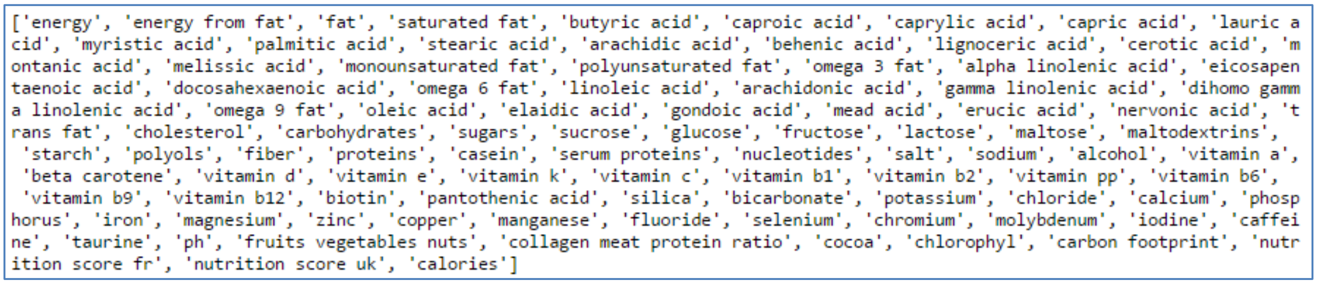
The FDA recommends to Americans certain levels of daily nutrient intake, based on IOM research. Do the foods available in our supermarkets generally allow us to meet these recommended intake levels? Are there certain nutrients that we must work harder to consume than others? We start with the list of nutrients for which IOM RDAs are available[1] and the full list of nutrients for which our dataset offers a field.

Figure 1: IOM Nutrients used in Daily Recommendations



	0	1	2	3	4	5
1	Vitamin A	Biotin	Riboflavin	Calcium	Iron	Potassium
2	Vitamin C	Choline	Thiamin	Chromium	Magnesium	Selenium
3	Vitamin D	Folate	Vitamin B-6	Copper	Molybdenum	Zinc
4	Vitamin E	Niacin	Vitamin B-12	Fluoride	Manganese	x
5	Vitamin K	Pantothenic Acid	x	Iodine	Phosphorus	x

Figure 2: All nutrients available from World Food Facts

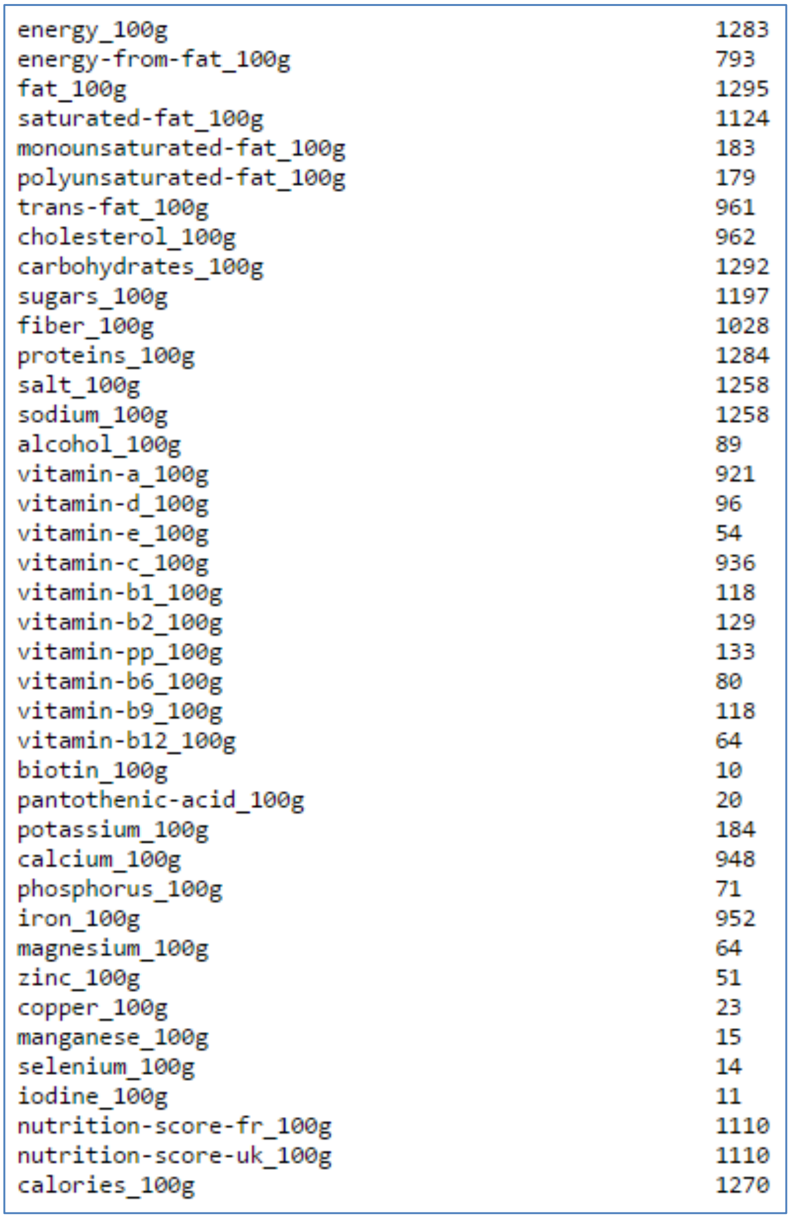


```
['energy', 'energy from fat', 'fat', 'saturated fat', 'butyric acid', 'caproic acid', 'caprylic acid', 'capric acid', 'lauric a
cid', 'myristic acid', 'palmitic acid', 'stearic acid', 'arachidic acid', 'behenic acid', 'lignoceric acid', 'cerotic acid', 'm
ontanic acid', 'melissic acid', 'monounsaturated fat', 'polyunsaturated fat', 'omega 3 fat', 'alpha linolenic acid', 'eicosapen
taenoic acid', 'docosahexaenoic acid', 'omega 6 fat', 'linoleic acid', 'arachidonic acid', 'gamma linolenic acid', 'dihomo gamm
a linolenic acid', 'omega 9 fat', 'oleic acid', 'elaidic acid', 'gondoic acid', 'mead acid', 'erucic acid', 'nervonic acid', 't
rans fat', 'cholesterol', 'carbohydrates', 'sugars', 'sucrose', 'glucose', 'fructose', 'lactose', 'maltose', 'maltodextrins',
 'starch', 'polyols', 'fiber', 'proteins', 'casein', 'serum proteins', 'nucleotides', 'salt', 'sodium', 'alcohol', 'vitamin a',
 'beta carotene', 'vitamin d', 'vitamin e', 'vitamin k', 'vitamin c', 'vitamin b1', 'vitamin b2', 'vitamin pp', 'vitamin b6',
 'vitamin b9', 'vitamin b12', 'biotin', 'pantothenic acid', 'silica', 'bicarbonate', 'potassium', 'chloride', 'calcium', 'phosp
horus', 'iron', 'magnesium', 'zinc', 'copper', 'manganese', 'fluoride', 'selenium', 'chromium', 'molybdenum', 'iodine', 'caffei
ne', 'taurine', 'ph', 'fruits vegetables nuts', 'collagen meat protein ratio', 'cocoa', 'chlorophyl', 'carbon footprint', 'nutr
ition score fr', 'nutrition score uk', 'calories']
```

Step One: Exploratory Analysis

Findings: Insufficient information is available at the point of purchase regarding the nutrient content of many products.

Figure 3: Frequency analysis of available nutrient information



```
energy_100g                                              1283
energy-from-fat_100g                                     793
fat_100g                                                 1295
saturated-fat_100g                                       1124
monounsaturated-fat_100g                                 183
polyunsaturated-fat_100g                                 179
trans-fat_100g                                           961
cholesterol_100g                                         962
carbohydrates_100g                                       1292
sugars_100g                                              1197
fiber_100g                                               1028
proteins_100g                                            1284
salt_100g                                                1258
sodium_100g                                              1258
alcohol_100g                                             89
vitamin-a_100g                                           921
vitamin-d_100g                                           96
vitamin-e_100g                                           54
vitamin-c_100g                                           936
vitamin-b1_100g                                          118
vitamin-b2_100g                                          129
vitamin-pp_100g                                          133
vitamin-b6_100g                                          80
vitamin-b9_100g                                          118
vitamin-b12_100g                                         64
biotin_100g                                              10
pantothenic-acid_100g                                    20
potassium_100g                                           184
calcium_100g                                             948
phosphorus_100g                                          71
iron_100g                                                952
magnesium_100g                                           64
zinc_100g                                                51
copper_100g                                              23
manganese_100g                                           15
selenium_100g                                            14
iodine_100g                                              11
nutrition-score-fr_100g                                  1110
nutrition-score-uk_100g                                  1110
calories_100g                                            1270
```

Many of the 150 data fields were sparsely populated in our subset. FDA labelling guidelines[2] only require reporting of select micronutrients, and leave most essential vitamins and minerals and all micronutrients under the umbrella of voluntary declaration[3]. Voluntary declarations are most commonly made on foods that present themselves as "healthier"[4]. We cannot conclude that foods that lack the declaration of these nutrients lack the nutrients themselves. Instead, we will adjust the list of micronutrients we seek to study from the full suite of 27 in the database, to those for which our dataset contains a meaningful number of entries (4 nutrients). We will likewise adjust our product list (2,800 items) to those that are well-described (900 items): Vitamin A, Vitamin C, iron and calcium. The new set, while smaller, was deemed sufficiently large to proceed.

Step Two: Tidy remaining data and verify

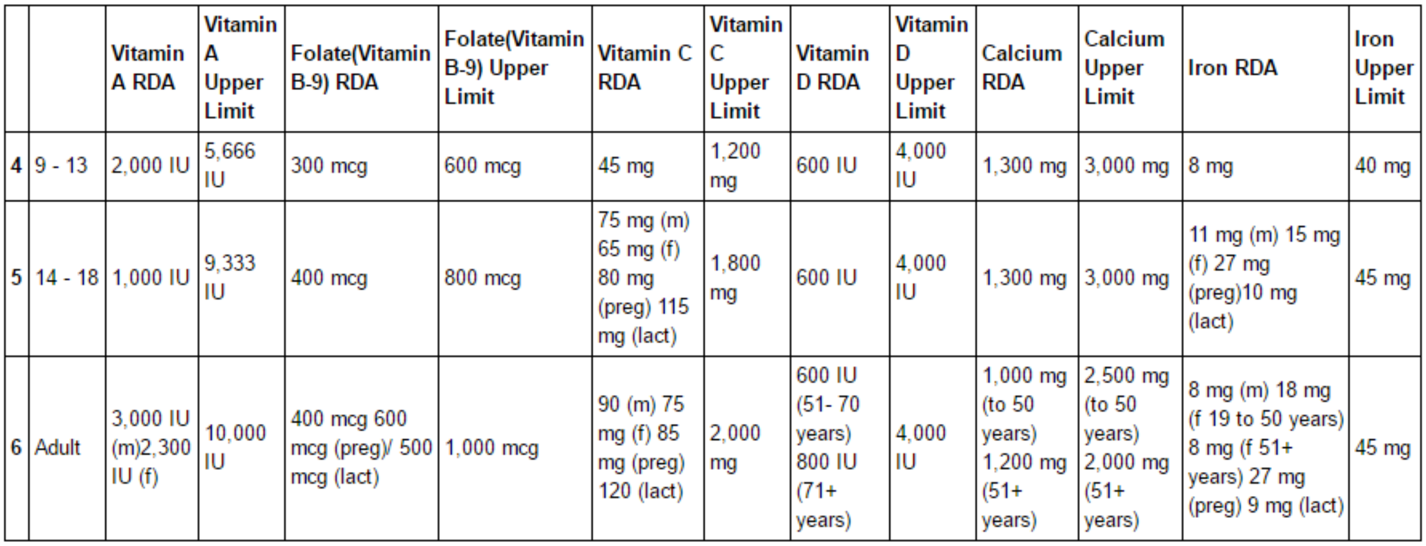
Findings: Nutrition information is provided in a variety of ways and wants for standardization.

After structuring the data set to include just the products for which sufficient nutrient information was present, and narrowing our list of interesting nutrients likewise, we worked to validate the data. The energy_100g field was particularly troubling; this field did not correlate to the calories per 100 grams as we had previously thought. Calories are the main qualifier of food and health, so it must inform the foundation for subsequent analysis. We chose to rebuild the field as an algebraic expression of the macronutrients, as random sampling showed us the macronutrient fields were reliable. Calories are defined as:

$$kcal = 9\left(fat_{gr}\right) + \ 4(carbohydrate_{gr} + protein_{gr})$$

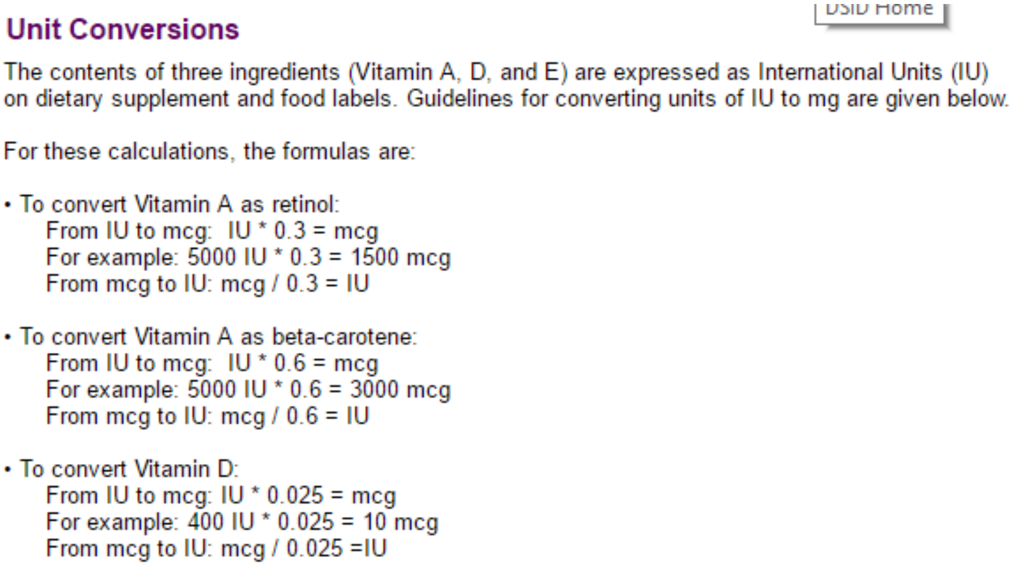
We also needed to verify that units were consistent among our data and between our working data sets (the product sample and the IOM table). Documentation included with the data had the units listed for the micronutrients as grams. However, the recommended daily intake of micronutrients is not measured in grams, but in micrograms, milligrams, and UI's (in the case of vitamin A). Vitamin A is especially tricky, as the daily recommendations for it change depending on the form. Vitamin A can take the form of retinol (animal source) or β-carotene (vegetable source; the *carot-* in carotene is a reference to the orange color of carrots). For instance, if a person consumes 5000 UI of vitamin A, this is either 1500 mcg of retinol or 3000 mcg of β-carotene. We converted the nutrients in two ways; as a global statistic and as a daily amount based upon a random day set we built.

Figure 4: Raw daily intake data



		Vitamin A RDA	Vitamin A Upper Limit	Folate(Vitamin B-9) RDA	Folate(Vitamin B-9) Upper Limit	Vitamin C RDA	Vitamin C Upper Limit	Vitamin D RDA	Vitamin D Upper Limit	Calcium RDA	Calcium Upper Limit	Iron RDA	Iron Upper Limit
4	9 - 13	2,000 IU	5,666 IU	300 mcg	600 mcg	45 mg	1,200 mg	600 IU	4,000 IU	1,300 mg	3,000 mg	8 mg	40 mg
5	14 - 18	1,000 IU	9,333 IU	400 mcg	800 mcg	75 mg (m) 65 mg (f) 80 mg (preg) 115 mg (lact)	1,800 mg	600 IU	4,000 IU	1,300 mg	3,000 mg	11 mg (m) 15 mg (f) 27 mg (preg)10 mg (lact)	45 mg
6	Adult	3,000 IU (m)2,300 IU (f)	10,000 IU	400 mcg 600 mcg (preg)/ 500 mcg (lact)	1,000 mcg	90 (m) 75 mg (f) 85 mg (preg) 120 (lact)	2,000 mg	600 IU (51- 70 years) 800 IU (71+ years)	4,000 IU	1,000 mg (to 50 years) 1,200 mg (51+ years)	2,500 mg (to 50 years) 2,000 mg (51+ years)	8 mg (m) 18 mg (f 19 to 50 years) 8 mg (f 51+ years) 27 mg (preg) 9 mg (lact)	45 mg

Figure 5: Conversion from IU to mcg



Unit Conversions

The contents of three ingredients (Vitamin A, D, and E) are expressed as International Units (IU) on dietary supplement and food labels. Guidelines for converting units of IU to mg are given below.

For these calculations, the formulas are:

- To convert Vitamin A as retinol:
  From IU to mcg: IU * 0.3 = mcg
  For example: 5000 IU * 0.3 = 1500 mcg
  From mcg to IU: mcg / 0.3 = IU

- To convert Vitamin A as beta-carotene:
  From IU to mcg: IU * 0.6 = mcg
  For example: 5000 IU * 0.6 = 3000 mcg
  From mcg to IU: mcg / 0.6 = IU

- To convert Vitamin D:
  From IU to mcg: IU * 0.025 = mcg
  For example: 400 IU * 0.025 = 10 mcg
  From mcg to IU: mcg / 0.025 =IU

Step Three: *Build an Average Daily Intake Profile*

Findings: An average sampling of these products provides sufficient vitamins and minerals.

A) *Global Statistics*

From the groomed data for the smaller set, we took a simple average on each remaining field. This represents the average content of that nutrient, among the products in our set. From this, we can build an average composite daily intake representation. Starting with calorie consumption recommendations (2500 for an adult male), and using our new calories_100g field, we determined that a person could consume 862.7g of food from our sample set, each day. This amount of food provided quantities of nutrients that were roughly in-line with recommended levels on all counts.

For the global analysis, we calculated that in order to meet the daily recommended intake of calories from this list, an individual must consume 867 grams of food per day, equating about 8.67 items from the list. Given the calories needs were met, we found the following about micronutrients:

- vitamin a consumed: 1038.28 mcg, recommended: 900 mcg
- vitamin c consumed: 51.50 mg recommended: 90 mg
- calcium consumed 788.93 mg recommended 1,000 mg
- iron consumed 20.52 mg recommended 8 mg

Given that a person eats 867 grams per day from this list, they are able to meet their vitamin A and their iron intakes, but not their vitamin C or calcium intake.

B) *Sampling Statistics*

Our dataset is admittedly incomplete. The average grocery store contains over 30,000 products to choose from, and our set barely contains 1,000 products. In addition, the fields themselves are not always complete. Thus, we wanted to use the statistical method of bootstrapping to build another model for parity. Bootstrapping was a perfect methodology for this set, as it assumes incomplete data gathering. We used bootstrapping to construct a set of 10,000 days and compared it to the global statistics.

While researching, we discovered that the average person consumes around 4 pounds or 2040 grams of food per day. We know that each entry in the dataset has a nutrient profile based upon 100 gram increments, so we started to select 20 items randomly using bootstrapping methods, and began calculations. We quickly halved our food intake assumption, as the calories were double what a person could reasonably eat in a day. Thus our random days were built upon 10 randomly selected items per day. After building our dataset, we plot the calorie data. We want to verify that the central limit theorem holds; and that the data is normally distributed around a center.

Figure 6: histogram of calories

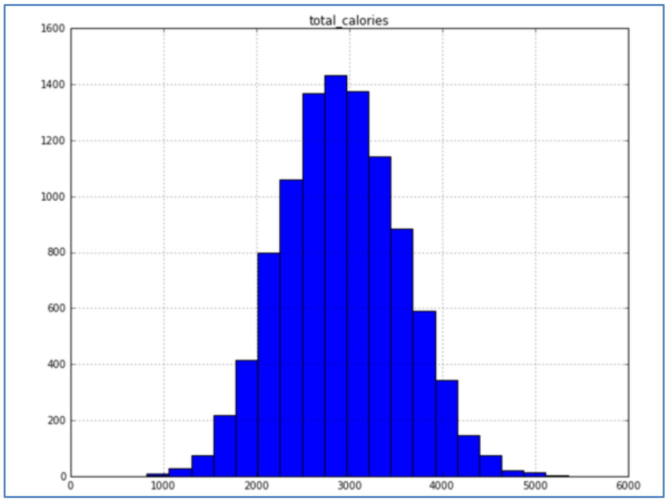


Figure 7: box and whisker plot of calorie distribution

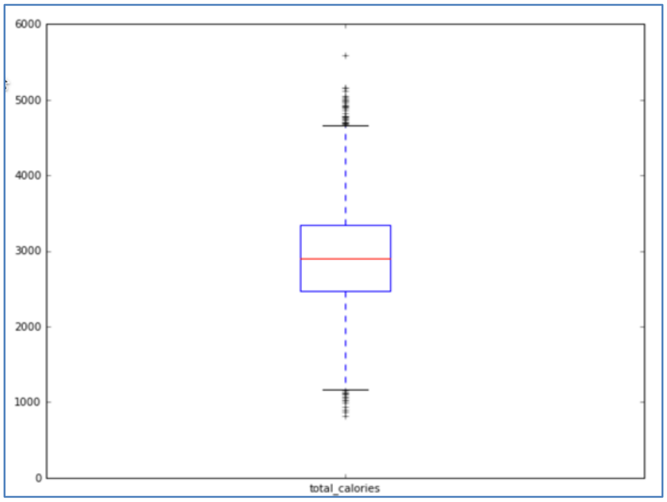
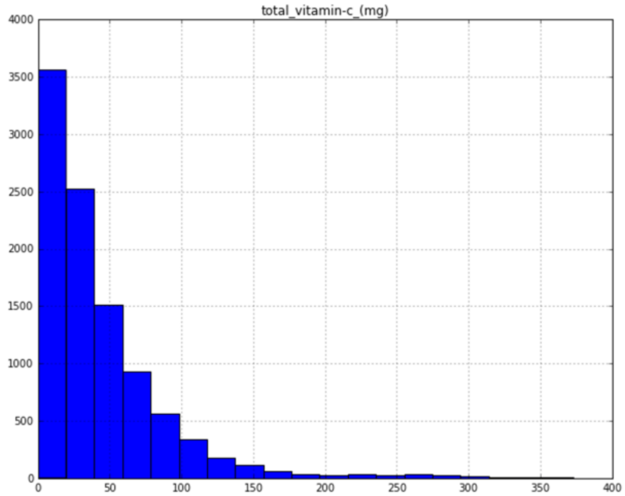


Figure 5: example of sample skew



Our calorie data is indeed normally distributed around a center. Calculating the skew of this set using SciPy's skew function gives us a skew of 0.10, which means this set is nearly normally distributed. These random days therefore can be used in the same way the global statistic was. These days represent more calories than the recommendation of 2,000 calories, but this is not unexpected. This set is biased against produced, and Americans are often criticized for eating too much

For our micronutrients, this set gives a slightly more conservative number:
vitamin a consumed: 868 mcg
recommended: 900 mcg
vitamin c consumed: 42 mg
recommended: 90 mg
calcium consumed 665 mg
recommended 1,000 mg
iron consumed 17.42mg
recommended 8 mg

Why is this? Upon graphing the data, we can see that the micronutrients are extremely skewed and do not follow a normal distribution. In addition, this set includes partially incomplete data; only the macronutrient fields were considered when creating this set. Thus, many of the random days have incomplete micronutrient data.

Question 2: Do these food contain any controversial or harmful ingredients that could cause health problems ?

One of the questions that we look into in this study is the ingredient lists. In this study, we decided to choose partially hydrogenated oil (PHO) and High fructose corn syrup(HFCS) and its presence in the packaged food in the US.

Hydrogenation, complete or partial, is a chemical process in which hydrogen is added to liquid oils to turn them into a solid form. PHOs, the primary source of industrially produced trans fat, are found in many popular processed foods, like baked goods and frozen foods that Americans use to feed their families to increase the shelf life.

FDA required all the packaged food in the US to list "trans fat" under nutrition label, and they recommended consumers to consider the amounts of saturated fat and trans fat, and choose the product that has the lowest amounts. Since trans fat that come from PHO has been linked to an increased risk of coronary heart disease by contributing to the buildup of plaque inside the arteries that may cause a heart attack, and it has been commonly used in the food industry including snack food, beverages, refrigerated food, bakeries etc., it would be interesting to analyze the percentage of packaged food that contain the PHO. In this study, keywords "hydrogenated" would be used as search tool.

High fructose corn syrup (HFCS) is derived from corn starch that has broken down with added enzyme to make it cheaper sweeter. HFCS is categorized by FDA as generally recognized as safe compare to other sweeteners, however, it has been controversial regarding the health risk of consuming HFCS on regular basis, and HFCS is not considered as consumer friendly ingredients. In this study, we use use keywords "high fructose" and "corn syrup" to look for the presence of high fructose corn syrup and corn syrup in the packaged food.

First step: Tidying the data and verify the information.

Similar with Question 1, the first step for this analysis is to gather the ingredient information and narrowing our list to the valid ingredients by substituting and filtering the invalid "NaN" with "X".

Second step:

Search for the keywords and create a function that reads the ingredients data and flags for keywords. In this function, any keyword could be used to search the keywords with flags. Flag array was created with 0 means no keyword and 1 contains keyword.

Third step:

Set up various counter and compute the statistics for this study.

Question 2 Findings:

1. Total number of valid ingredients (without NANs) in this database is 913.
2. The frequency of the ingredients that contain hydrogenated or hydrogenated oil is 68 out of 913, which is 7.45%. The frequency of the high fructose is 89 out of 913, which is 9.75%. The frequency of the corn syrup is 158 out of 913, which is 17.31%. The percentage of recipes containing either 'corn syrup', 'hydrogenated', 'high fructose' is 22.35%.

Question 3: Are these food considered as "healthy" or "unhealthy" ?

Next, we would like to determine if the food in this database could be considered as healthy or unhealthy food based on their nutritional content.

Nutrient profiling is the science of classifying or ranking foods according to their nutritional composition for reasons related to preventing disease and promoting health. Though nutrient profiling does not address all aspects of nutrition, diet and health it is a helpful tool to use in conjunction with interventions aimed at improving diets.

The United States FDA does not have a current nutritional profiling tool, we are using UK UK Food Standards Agency as reference in this study to categorize and determine the data in this study.

Overall score = (total 'A' points) minus (fiber points + fruit, veg and nuts points only) [i.e. not allowed to score points for protein]

A food is classified as 'less healthy' where it scores 4 points or more.
A drink is classified as 'less healthy' where it scores 1 point or more.

First step: Tidying the data and verify the information.

Similar with Question 1 and 2, the first step for this analysis is to gather the nutrition score information and narrowing our list to the valid ingredients by substituting and filtering the invalid "NaN" with "X".

Second step:

After filtering out the invalid data, next we would like to determine if the food is "healthy" or "unhealthy". We would calculate the total nutrition score in this database; the percentage of the score that is equal or larger than 4; the average number of the UK nutrition score.

Third step:

After determining the nutrition score in this study, we could hypothesize this database as a random selection of a larger database, and we could calculate the student t-score and p-value by using Scipy.

Question 3 Findings:

1. The total number of UK nutrition score or profiling in this database is 1110, the percentage of the UK nutrition score that is equal or larger than 4 is 59.19%, the average number of the UK nutrition score in this database is 8.78.
2. We assume that the database represents a random sample of packaged food in the US. We hypothesize that the average UK nutrition score for all the packaged food in the U.S. is larger than 4, which means the foods are classified as ‘less healthy’. The sample's standard deviation in this database is 9.42. The t-score of the hypothesis that the average is larger then 4 is 16.91, for 1109 degrees of freedom. The p value of the hypothesis test that the UK nutrition score equal or larger than 4 is 1.6794485891e-57 , hence, the average nutrition score is significantly larger than 4 at a significant level of 5%.
3. In conclusion, most of the nutrition score in this study would be considered as "unhealthy" by the UK Food Standards Agency.

Biases & Limitations:¶

We sized down our data set at several points. Our final analysis was based on a 900 X 15 dataframe. Had more of the fields been populated, or had our starting set been larger (French products for example in Open Food Facts number greater than 56,000) there may have been more room for nuanced analysis.

The data set is built on information from each food's packaging (versus, for example, laboratory analysis). As such it reflects and is limited by industry standards and regulations. The US FDA requires reporting of only a few choice nutrients. Voluntary reporting of non-statutory nutrients is limited. Our analysis, therefore, cannot be considered an examination of the true content and nutritional value of these foods, but rather the information that a consumer of the food could reasonably discern. In all likelihood, a diet of these foods would be more robust and nutritionally satisfying than we can with certainty say here.

Likewise, a set of packaged foods will naturally deselect for some important food categories. Fruits, vegetables, bulk foods (nuts, beans, grains) do not appear in our set[5]. The role of these products in an average consumer's diet and more importantly their nutrient intake, is categorically understated in this study.

References:

1. https://www.consumerlab.com/RDAs/
2. https://www.federalregister.gov/articles/2016/05/27/2016-11867/food-labeling-revision-of-the-nutrition-and-supplement-facts-labels#h-31
3. FDA considers required label inclusion for "non-statutory nutrients...for which there is an independent relationship between the nutrient and risk of chronic disease, health-related condition, or physiological endpoint."
4. http://world.openfoodfacts.org/product/0082592720153
5. http://www.fda.gov/Food/IngredientsPackagingLabeling/FoodAdditivesIngredients/ucm449162.htm
6. https://www.federalregister.gov/articles/2015/06/17/2015-14883/determinations-partially-hydrogenatedoils
7. http://www.fda.gov/Food/IngredientsPackagingLabeling/FoodAdditivesIngredients/ucm324856.htm
8. http://www.food.gov.uk/sites/default/files/multimedia/pdfs/techguidenutprofiling.pdf
9. https://www.food.gov.uk/northern-ireland/nutritionni/niyoungpeople/nutlab/nutprofmod
10. http://www.who.int/nutrition/topics/profiling/en/