**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.

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

**Question 1: Are these foods nutritious?**

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 available1 and the full list of nutrients for which our dataset offers a field.

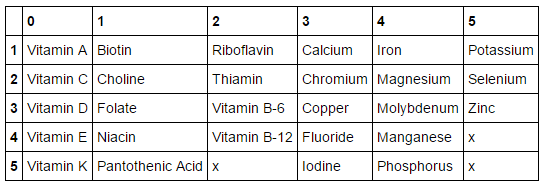


Fig 1: Nutrients for which IOM makes daily intake recommendations

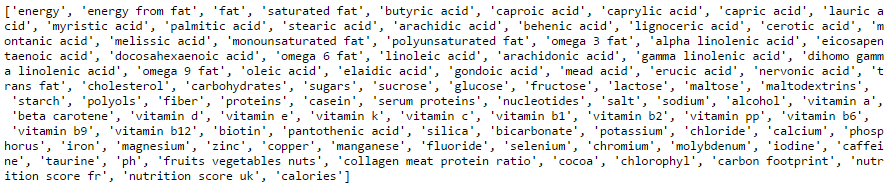
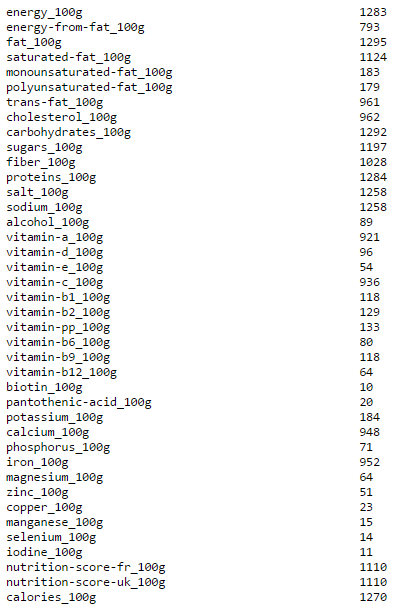


Fig 2: Full list of nutrients in our data set

***Step One:* Exploratory Analysis**

***Finding:*Insufficient information is available at the point of purchase, regarding the nutrient content of many of these products.**

Many of the 150 data fields were sparsely populated in our subset. FDA labelling guidelines2 only require reporting of select macronutrients, and leave most essential vitamins and minerals and all micronutrients under the umbrella of voluntary declaration3. 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 nutrients 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). The new set, while smaller, was deemed sufficiently large to proceed.

We are left to study: Vitamin A, Vitamin C, Calcium, and Iron.

Fig 3: Nutrients list and frequency count

#### Step Two: Verify Important Data

#### Finding: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; as this was to form the foundation for subsequent analysis, we chose to rebuild the field using data that we were more confident was reliable. We also needed to transform some of the data so that units were consistent among our data and between our working data sets (the product sample and the IOM table).

#### Step Three: Build an Averge Daily Intake Profile

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

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.

### Biases & Limitations:[¶](http://localhost:8888/notebooks/python_course/Practice%20Code%20%26%20Problems/CP_Project_Question_One.ipynb" \l "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 set5. The role of these products in an average consumer's diet and more importantly their nutrient intake, is categorically understated in this study.

### Question One Conclusion:

Packaged foods available in the US do not provide a robust source of declared nutrients. Vitamin B is particularly deficient.

1. source: <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>
2. 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."
3. <http://world.openfoodfacts.org/product/0082592720153>
4. Chart that includes "vegetable" type product categories & their counts