Open Food Facts

Link: <https://www.kaggle.com/datasets/openfoodfacts/world-food-facts>

Texts that can be inserted in to the notebook: - Xiaoyun Yu

Coconut oils are healthy.

Hydrogenate fats are trans fats, are unhealthy.

Brominated vegetable oil is unhealthy (horrible). Brominated oil is to make drinks thicker.

Bromine (added to breads) is banned in US since 1991 and China.

Fructose is for our ancestors to gain weights and fats and to decrease brain activity before the long winter [1].

Assess the dataset:

**Quality: Is the data reliable? On what basis do you make that judgement?**

Yes. The dataset is provided directly by the owner called “open food facts”, and Kaggle team. There are 542 codes already discussing about this dataset. And it even has an AI repository. We see this is a popular dataset, with over 71k downloads.

**Detail: How much detail is there? Is the information helpful?**

There are tremendous details. This dataset includes 356027 rows × 163 columns, although some columns are empty or sparse that should be discarded. Each food entry also has a URL which is a webpage containing tidy visualizations for its brief info. The columns contain not only the food’s ingredients in full detail, but also the import country and city, shop, and creator. The countries mainly contain European and American countries.

**Documentation: How clear is it what the data means? Where was the documentation? Was it easy to find?**

Is lacks of source/provenance, and column descriptions. This is one of its drawbacks. However, we can infer what each column means just by its names. Also, Kaggle’s preview functionality is powerful for us to have an overview for each column’s values. It says that “the data is a combination of crowdsourced information gathered via mobile phones, producer data, and data extracted from the first two sources using various machine learning, OCR, and regex techniques.” (<https://www.kaggle.com/datasets/openfoodfacts/world-food-facts>)

**Interrelation: Would it be useful to connect to other data sets? Which? How easy would that be?**

It is difficult to find another related dataset, since this dataset is mainly based on each of the specific food names. It is rather difficult to find the specific food names. However, we may find another dataset with country-city entries to be combined with this dataset.

**Use: What could you use it for? What questions would you like to ask the dataset but can't? What's missing?**

We can:

(1) Do classification using all necessary columns against “main\_category\_en” column.

(2) Use “ingredients\_text” to classify “calcium\_100g” column value into: low, med, high.

(3) Use other columns to predict “nutrition-score-uk\_100g” column value.

(4) Use other columns to predict “energy\_100g” column value.

(5) Do classification using all necessary columns against “countries\_tags” column.

(6) Find any relationships between the countries and “nutrition-score-uk\_100g” column.

(7) Environmental Threat Analysis: Use an XPath approach to extract a Boolean flag indicating whether a product poses a species threat.

(8) Nutritional Trends & EDA: Investigate the correlation between high sugar content, additive levels, and nutrition scores across countries. (Inference)

We can’t:

(1) Analyze the Eco-score directly from the table, since that score is a missing column.

**Discoverability: How easy was it to find open data in your chosen domain? Where did you go? Were there many alternatives?**

For food data, there are tons of food images classification data in Kaggle. However there are few with plain data.

**Terms of use / license:**

[Database: Open Database, Contents: Database Contents](http://opendatacommons.org/licenses/dbcl/1.0/)

**Interests:**

**Difficulties:**

(1) Countries and cities are ambiguous. In some columns, they are mixed, separated with commas; in other columns, there are only countries, or there are codes such as “brignemont-haute-garonne-france” that are hard to understand. Since there are non-UDF characters, and multiple natural languages, difficulty of data validation and entity resolution (coreference). For example, “en:GB”, “United Kingdom” and “Royaume-Uni” are the same thing.

(2) There are many columns with collinearity. For example, similar columns: “ingredients\_from\_palm\_oil\_tags”, “ingredients\_from\_palm\_oil”, “ingredients\_from\_palm\_oil\_n”.

(3) “ingredients\_text” column is hard to expand, since there are parentheses and commas (layered structure), and the ingredient’s name may be different for the same ingredient (entity resolution).

(4) Since it’s lack of metadata, it’s hard to infer for some technical terms, such as “E307c - Tocopherol” and “fr:filet-de-boeuf”.

(5) Imbalanced dataset. For example, there are predominant foods which are from the United States.

(6) Because of the different serving\_size for each entry, we can’t directly compare some of the numerical values of each entry.

To solve the issues of some anomalies that may occur during our management of the database, and to maintain ACID properties, we need to normalize the database. Normal anomalies may include: update anomaly, insertion anomaly, deletion anomaly, modification anomaly etc. Normal forms define sets of attributes that ensure relations are properly structured. They are progressive, with each successive normal form imposing increasingly stringent requirements.

|  |  |
| --- | --- |
| 1NF |  |
| ↓ | eliminate the partial functional dependencies of non-prime attributes towards keys |
| 2NF |  |
| ↓ | eliminate the transitive functional dependencies of non-prime attributes towards keys |
| 3NF |  |
| ↓ | eliminate the partial transitive functional dependencies of prime attributes towards keys |
| BCNF |  |

The 1st normal form emphasizes on **Atomicity**. “A relation is in first normal form (1NF) if and only if all the domains in which its attributes are defined contain scalar values only.” (Lewis, Normalisation, p. 113) For this dataset to be in the 1st normal form, we should expand each of the multiple-value element to make each cell only containing one element. The “code” column shouldn’t be set as the unique ID, since there are only 355839 values in that column, and we have 356027 rows. We can set the “product\_id” as the primary key for each row. However, for the “states” columns, we can just leave multiple-value elements there (or separate these into another table), since the “states” is a summary for all the other info.

States:

|  |  |  |  |
| --- | --- | --- | --- |
| states | States\_tags | states\_en | product\_id |
|  |  |  |  |

**Not satisfying 2NF: Some attributes are not determined by the primary key**. For example, creator is not determined by product\_id.

A relation is in second normal form (2NF) if and only if: 1. it is in 1NF; and 2. **every non-key attribute is irreducibly dependent on the primary key** (Lewis, Normalisation, p. 113). It deletes the partial functional dependency of non-prime attributes towards keys.

We need to create one table for the creator, and use product\_id as a foreign key.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Creator\_id | creator | created\_t | created\_datetime | last\_modified\_t | last\_modified\_datetime | product\_id |
|  |  |  |  |  |  |  |

For packaging and brands, we can temporarily see these as properties for each food product.

For city and country, we should create separate tables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| id | origin | origin\_tag | manufacturing\_place | manufacturing\_place\_tag | product\_id |
|  |  |  |  |  |  |

City-country:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| id | city | city\_tag | purchase\_place | country | country\_tag | country\_en | product\_id |
|  |  |  |  |  |  |  |  |

Country-food:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | country | country\_tag | country\_en | product\_id |
|  |  |  |  |  |

Categories:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | category | categories\_tag | category\_en | product\_id |
|  |  |  |  |  |
|  |  |  |  |  |

Stores:

|  |  |  |
| --- | --- | --- |
| store\_id | store | product\_id |
|  |  |  |

States (leave as what it is):

|  |  |  |  |
| --- | --- | --- | --- |
| product\_id | states | States\_tags | states\_en |
|  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | emb\_code | emb\_codes\_tag | first\_packaging\_code\_geo | product\_id |
|  |  |  |  |  |

Introduction

Think of your favorite food! Is it crispy Lay’s chips? Or cold Coca-Cola for which a man is drinking freely during a hot summer? I bet it’s not Chinese cabbage. However, most of the packaged foods are not as nutritious as the natural foods, and even more, many manufactured foods have misleading appealing packages or advertisements, or, some may cheat you: 75% honey is really not made by natural honey – they are made by high fructose corn syrup, added by flavoring, thickener, and color [1]; some may write “No Trans Fat” on their package, but actually writes “hydrogenated soybean oil” (trans-fat) in the ingredients.



In the Nov 2011 “National Convention for the Dietitians”, we can see a lady dressed in red, representing heart protection, and she is sponsored by Coca-Cola. Dr. Schmidt says, it implies that the food industries can control the American Dietitians Association, teaching the dietitians how to speak to the patients [1]. It means that, there may be conspiracies and undiscovered truths behind the ingredients for food, and our illness. “Food has three functionalities: (1) provide energy; (2) sustain life; (3) repair tissues. ” [1] Therefore, if the packaged foods cannot have these three functionalities, they can be seen as garbage.

But don’t be so pessimistic. Luckily, there is a fair rule for all foods – they must have ingredient texts and nutrition list! Therefore, we have a compelling reason to dive into the ingredients and origins of the packaged foods, to see how much nutrition they really have, and how healthy or unhealthy they could be. Moreover, there may be interesting facts among countries.

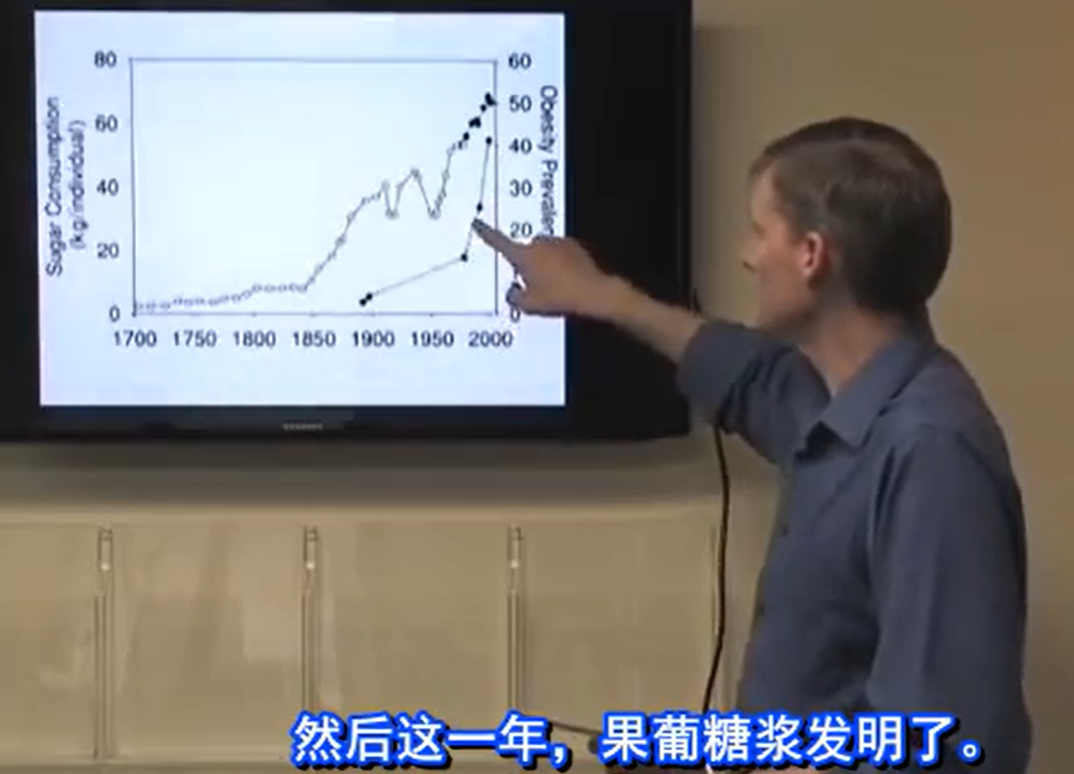
We have found the “Open Food Facts” database from Kaggle.com. This database is created officially by the owner called “open food facts”, and Kaggle team. Although it is a crowdsource project, the data are fairly complete and valuable. These foods are mainly packaged foods, with rather longer quality-guarantee periods (than the natural foods), which can maintain the transportations among cities and even countries.

**Hypothesis testing:** Dr. Schmidt has introduced a case, where his friends had a vacation to Sweden and Iceland and their illness and fatigues went away, which he summarized as the high-quality of their foods compared to American foods. Is this hypothesis true? He emphasizes that, when the sugar and carbohydrates go down, the food’s quality go up. So, we can simplify the hypothesis as **H1: the foods from Sweden and Iceland have less sugar and carbohydrate, compared to American foods. Our H0: the foods from Sweden and Iceland have as much sugar and carbohydrate as American foods**

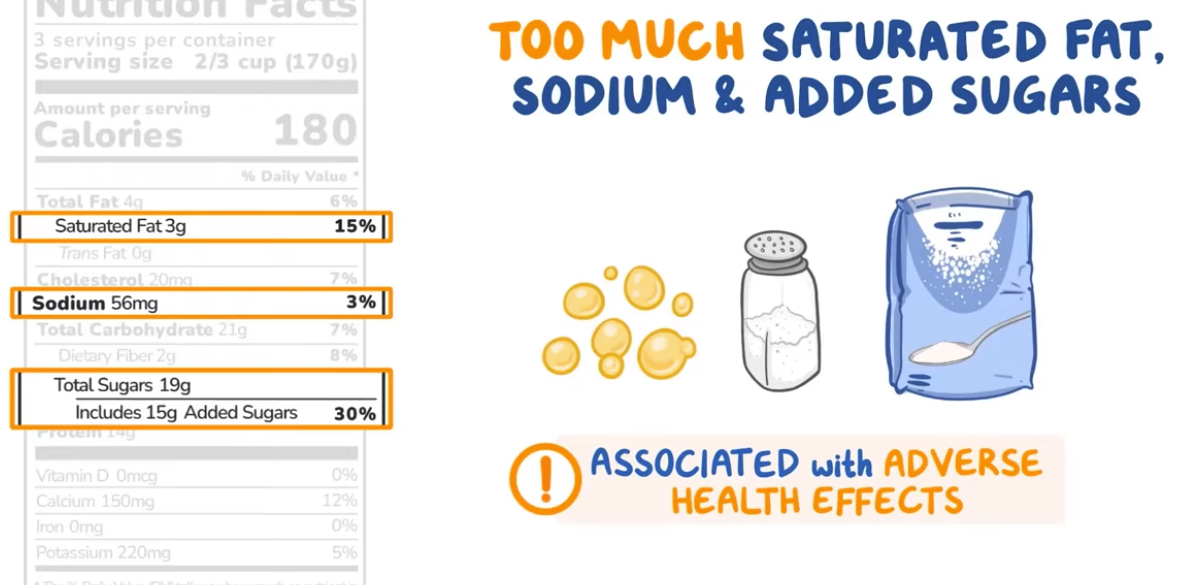
**Hypothesis testing:** Is it true that low-fat foods generally have more carbohydrate to add to the flavor?

During the past 60 years, Americans are more and more sick, which is not caused by smoking or alcohols (because these rates are dropping down), but is caused by high-carbohydrate diet (mulnutrition) [1].

Discussions:



High fructose corn syrup (果葡糖浆): According to Dr. Schmidt, 50% High fructose corn syrup can detect toxic mercury. Besides, during the year when High fructose corn syrup was invented, obesity rate went up [1].



Discussions:

Many additives are not directly added to the food, but are sometimes added to the preprocessing of the food, which can be avoided to be written to the ingredients text. For example, Olacta (乳化剂) is not directly added to the food, but instead is added to the dough, which can strength the dough but is bad to our health.

MSG (味精/谷氨酸钠) can make people fatter. Unfortunately there are 60+ names for MSG, which is difficult for us to do entity resolution (coreference resolution), since there are hidden amount of MSG inside of other gradients’ names, such as Gelatin, Sodium Caseinate, or even Calcium [2].

Is the food name actually a good source for indication of nutrition level? No, because most of them are exaggerating! Food industries’ priority is not to make you healthy. It is to make money. If we have time, we can further look at the foods’ names – compared with the ingredients, which foods are cheating us? In other words, which foods do not have the ingredients or nutrition level, but are named by that key ingredient? Which countries are they from? And which countries’ people often buy these foods?

The food’s nutrition should often not be measured by quantity, but by quality instead. For example, salts are salts, but organic French grey sea salt is more nutritious and healthier than fine salts, where the former has minerals, but the latter only has salt (and even lists minerals as impurities) [1].

Beet sugar and refined sugar are completely different [1]. Dr. Schmidt says that Iceland people eat beet sugar every day, but none of them are overweight. It is difficult to differentiate these two kinds of sugar, since they are both called “sugar” in the nutrition facts label – unless you explore the ingredient texts, and that’s another amount of work.

Again, quality wins. Healthy and happy free-ranging chickens and cows are better than those are only fed inside of the cage or fence. However this info is not indicated in the nutrition facts list, nor in the ingredient texts.

Another hidden issue is that, many plants have pesticides, which are not indicated in the ingredient texts either because they never admit, or because there is not a compulsory rule.

The “calory” column is based on one serving size, and that serving size is different among foods.

Artificial sweeteners such as Aspartame can lead to zero sugar in the ingredients, but Aspartame can make you more crave for the unhealthy food!

It’s even more interesting if the machine can identify the fake and exaggerated texts on the package, versus their actual ingredients.

Citations (IEEE style)

[1] Schmidt, Darren. “Food Freak Show-疾病到底从哪来？美国营养师解密食品巨头的阴谋(Where do diseases come from? American nutritionists reveal the conspiracy of food giants)\_bilibili,” bilibili, <https://www.bilibili.com/video/BV1sb411q7YZ/> (accessed Nov. 10, 2024).

[2] “MSG Hidden names,” MSGmyth, <https://www.msgmyth.com/hidden-names> (accessed Nov. 10, 2024).

[3] Osmosis from Elsevier, “How to read a nutrition facts label,” YouTube, <https://www.youtube.com/watch?v=hzwnwLgmFOQ> (accessed Nov. 10, 2024).

[4] Doctor Mike, “Top 10 Misleading Food Label Claims | Nutrition Labels BUSTED!!!,” YouTube, <https://www.youtube.com/watch?v=JXDo-73uaAI> (accessed Nov. 10, 2024).