# Sobre los datos.

Utilizaremos los datos de “Open Food Facts”[[1]](#footnote-1), que es una base de datos de productos alimenticios con ingredientes, alérgenos, información nutricional y toda la información que podemos encontrar en las etiquetas de los productos. Esta información de los productos alimenticios (fotos, ingredientes, información nutricional, etc.) se recopila de forma colaborativa y se pone a disposición de todos y para todos los usos en una base de datos abierta y gratuita. Para permitir que los usuarios verifiquen los datos, se invita a los contribuyentes a subir fotos de etiquetas y empaques que muestren los datos.

La base de datos Open Food Facts está disponible bajo la licencia “Open Data Commons”[[2]](#footnote-2), que es un acuerdo de licencia destinado a permitir que los usuarios compartan, modifiquen y usen libremente esta base de datos mientras mantienen esta misma libertad para los demás. Autorizan el uso y la reproducción del contenido para todos los fines, incluido el uso comercial, bajo ciertas condiciones, especialmente la atribución y la distribución en las mismas condiciones de los trabajos derivados.

Las imágenes de los productos están disponibles bajo la licencia “Creative Commons Attribution ShareAlike”[[3]](#footnote-3). Pueden contener elementos gráficos sujetos a derechos de autor u otros derechos, que en algunos casos pueden ser reproducidos (derechos de cita o uso justo). La licencia “Creative Commons Attribution ShareAlike” para imágenes solo cubre la foto en sí y los derechos de los colaboradores que tomaron la foto. Pueden aplicarse otros derechos de terceros, como: derechos de autor del diseño del producto y los elementos gráficos que contiene (ilustraciones, imágenes, etc.), derechos de imagen de las personas (por ejemplo, celebridades) que aparecen en el embalaje, derechos de marca registrada, etc.

# Trabajo de datos.

Para trabajar los datos, descargamos todos los datos en formato CSV sin procesar desde la pagina de Open Food Facts (<https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv>), este archivo tiene un tamaño de 7,54 GB (8.104.795.848 bytes).

Para ello utilizamos Google Colab para que todos los integrantes tuvieran acceso y participaran del procesamiento de la información.

La lectura de los datos se realizó de la siguiente manera:

Texto, Aplicación

Descripción generada automáticamente

De aquí podemos rescatar en primera instancia que tenemos un total de 2.850.798 de productos en esta base de datos, de los cuales 1.714.180 son distintos (en cuanto al nombre).

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Además, observamos que las columnas y su respectivo tipo de datos son las siguientes:

+-----------------------------------------------------+---------------+----------------+

|Columna |Tipo de datos |Tipo de variable|

+-----------------------------------------------------+---------------+----------------+

|code |DoubleType() |Cuantitativa |

|url |StringType() |Cualitativa |

|creator |StringType() |Cualitativa |

|created\_t |IntegerType() |Cuantitativa |

|created\_datetime |TimestampType()|Cuantitativa |

|last\_modified\_t |IntegerType() |Cuantitativa |

|last\_modified\_datetime |TimestampType()|Cuantitativa |

|last\_modified\_by |StringType() |Cualitativa |

|product\_name |StringType() |Cualitativa |

|abbreviated\_product\_name |StringType() |Cualitativa |

|generic\_name |StringType() |Cualitativa |

|quantity |StringType() |Cualitativa |

|packaging |StringType() |Cualitativa |

|packaging\_tags |StringType() |Cualitativa |

|packaging\_en |StringType() |Cualitativa |

|packaging\_text |StringType() |Cualitativa |

|brands |StringType() |Cualitativa |

|brands\_tags |StringType() |Cualitativa |

|categories |StringType() |Cualitativa |

|categories\_tags |StringType() |Cualitativa |

|categories\_en |StringType() |Cualitativa |

|origins |StringType() |Cualitativa |

|origins\_tags |StringType() |Cualitativa |

|origins\_en |StringType() |Cualitativa |

|manufacturing\_places |StringType() |Cualitativa |

|manufacturing\_places\_tags |StringType() |Cualitativa |

|labels |StringType() |Cualitativa |

|labels\_tags |StringType() |Cualitativa |

|labels\_en |StringType() |Cualitativa |

|emb\_codes |StringType() |Cualitativa |

|emb\_codes\_tags |StringType() |Cualitativa |

|first\_packaging\_code\_geo |StringType() |Cualitativa |

|cities |StringType() |Cualitativa |

|cities\_tags |StringType() |Cualitativa |

|purchase\_places |StringType() |Cualitativa |

|stores |StringType() |Cualitativa |

|countries |StringType() |Cualitativa |

|countries\_tags |StringType() |Cualitativa |

|countries\_en |StringType() |Cualitativa |

|ingredients\_text |StringType() |Cualitativa |

|ingredients\_tags |StringType() |Cualitativa |

|ingredients\_analysis\_tags |StringType() |Cualitativa |

|allergens |StringType() |Cualitativa |

|allergens\_en |StringType() |Cualitativa |

|traces |StringType() |Cualitativa |

|traces\_tags |StringType() |Cualitativa |

|traces\_en |StringType() |Cualitativa |

|serving\_size |StringType() |Cualitativa |

|serving\_quantity |DoubleType() |Cuantitativa |

|no\_nutrition\_data |StringType() |Cualitativa |

|additives\_n |IntegerType() |Cuantitativa |

|additives |StringType() |Cualitativa |

|additives\_tags |StringType() |Cualitativa |

|additives\_en |StringType() |Cualitativa |

|nutriscore\_score |IntegerType() |Cuantitativa |

|nutriscore\_grade |StringType() |Cualitativa |

|nova\_group |IntegerType() |Cuantitativa |

|pnns\_groups\_1 |StringType() |Cualitativa |

|pnns\_groups\_2 |StringType() |Cualitativa |

|food\_groups |StringType() |Cualitativa |

|food\_groups\_tags |StringType() |Cualitativa |

|food\_groups\_en |StringType() |Cualitativa |

|states |StringType() |Cualitativa |

|states\_tags |StringType() |Cualitativa |

|states\_en |StringType() |Cualitativa |

|brand\_owner |StringType() |Cualitativa |

|ecoscore\_score |DoubleType() |Cuantitativa |

|ecoscore\_grade |StringType() |Cualitativa |

|nutrient\_levels\_tags |StringType() |Cualitativa |

|product\_quantity |DoubleType() |Cuantitativa |

|owner |StringType() |Cualitativa |

|data\_quality\_errors\_tags |StringType() |Cualitativa |

|unique\_scans\_n |IntegerType() |Cuantitativa |

|popularity\_tags |StringType() |Cualitativa |

|completeness |DoubleType() |Cuantitativa |

|last\_image\_t |IntegerType() |Cuantitativa |

|last\_image\_datetime |TimestampType()|Cuantitativa |

|main\_category |StringType() |Cualitativa |

|main\_category\_en |StringType() |Cualitativa |

|image\_url |StringType() |Cualitativa |

|image\_small\_url |StringType() |Cualitativa |

|image\_ingredients\_url |StringType() |Cualitativa |

|image\_ingredients\_small\_url |StringType() |Cualitativa |

|image\_nutrition\_url |StringType() |Cualitativa |

|image\_nutrition\_small\_url |StringType() |Cualitativa |

|energy-kj\_100g |DoubleType() |Cuantitativa |

|energy-kcal\_100g |DoubleType() |Cuantitativa |

|energy\_100g |DoubleType() |Cuantitativa |

|energy-from-fat\_100g |DoubleType() |Cuantitativa |

|fat\_100g |DoubleType() |Cuantitativa |

|saturated-fat\_100g |DoubleType() |Cuantitativa |

|butyric-acid\_100g |DoubleType() |Cuantitativa |

|caproic-acid\_100g |DoubleType() |Cuantitativa |

|caprylic-acid\_100g |DoubleType() |Cuantitativa |

|capric-acid\_100g |DoubleType() |Cuantitativa |

|lauric-acid\_100g |DoubleType() |Cuantitativa |

|myristic-acid\_100g |DoubleType() |Cuantitativa |

|palmitic-acid\_100g |DoubleType() |Cuantitativa |

|stearic-acid\_100g |DoubleType() |Cuantitativa |

|arachidic-acid\_100g |DoubleType() |Cuantitativa |

|behenic-acid\_100g |DoubleType() |Cuantitativa |

|lignoceric-acid\_100g |DoubleType() |Cuantitativa |

|cerotic-acid\_100g |DoubleType() |Cuantitativa |

|montanic-acid\_100g |DoubleType() |Cuantitativa |

|melissic-acid\_100g |DoubleType() |Cuantitativa |

|unsaturated-fat\_100g |DoubleType() |Cuantitativa |

|monounsaturated-fat\_100g |DoubleType() |Cuantitativa |

|polyunsaturated-fat\_100g |DoubleType() |Cuantitativa |

|omega-3-fat\_100g |DoubleType() |Cuantitativa |

|alpha-linolenic-acid\_100g |DoubleType() |Cuantitativa |

|eicosapentaenoic-acid\_100g |DoubleType() |Cuantitativa |

|docosahexaenoic-acid\_100g |DoubleType() |Cuantitativa |

|omega-6-fat\_100g |DoubleType() |Cuantitativa |

|linoleic-acid\_100g |DoubleType() |Cuantitativa |

|arachidonic-acid\_100g |DoubleType() |Cuantitativa |

|gamma-linolenic-acid\_100g |DoubleType() |Cuantitativa |

|dihomo-gamma-linolenic-acid\_100g |DoubleType() |Cuantitativa |

|omega-9-fat\_100g |DoubleType() |Cuantitativa |

|oleic-acid\_100g |DoubleType() |Cuantitativa |

|elaidic-acid\_100g |DoubleType() |Cuantitativa |

|gondoic-acid\_100g |DoubleType() |Cuantitativa |

|mead-acid\_100g |DoubleType() |Cuantitativa |

|erucic-acid\_100g |DoubleType() |Cuantitativa |

|nervonic-acid\_100g |DoubleType() |Cuantitativa |

|trans-fat\_100g |DoubleType() |Cuantitativa |

|cholesterol\_100g |DoubleType() |Cuantitativa |

|carbohydrates\_100g |DoubleType() |Cuantitativa |

|sugars\_100g |DoubleType() |Cuantitativa |

|added-sugars\_100g |DoubleType() |Cuantitativa |

|sucrose\_100g |DoubleType() |Cuantitativa |

|glucose\_100g |DoubleType() |Cuantitativa |

|fructose\_100g |DoubleType() |Cuantitativa |

|lactose\_100g |DoubleType() |Cuantitativa |

|maltose\_100g |DoubleType() |Cuantitativa |

|maltodextrins\_100g |DoubleType() |Cuantitativa |

|starch\_100g |DoubleType() |Cuantitativa |

|polyols\_100g |DoubleType() |Cuantitativa |

|erythritol\_100g |DoubleType() |Cuantitativa |

|fiber\_100g |DoubleType() |Cuantitativa |

|soluble-fiber\_100g |DoubleType() |Cuantitativa |

|insoluble-fiber\_100g |DoubleType() |Cuantitativa |

|proteins\_100g |DoubleType() |Cuantitativa |

|casein\_100g |DoubleType() |Cuantitativa |

|serum-proteins\_100g |DoubleType() |Cuantitativa |

|nucleotides\_100g |DoubleType() |Cuantitativa |

|salt\_100g |DoubleType() |Cuantitativa |

|added-salt\_100g |DoubleType() |Cuantitativa |

|sodium\_100g |DoubleType() |Cuantitativa |

|alcohol\_100g |DoubleType() |Cuantitativa |

|vitamin-a\_100g |DoubleType() |Cuantitativa |

|beta-carotene\_100g |DoubleType() |Cuantitativa |

|vitamin-d\_100g |DoubleType() |Cuantitativa |

|vitamin-e\_100g |DoubleType() |Cuantitativa |

|vitamin-k\_100g |DoubleType() |Cuantitativa |

|vitamin-c\_100g |DoubleType() |Cuantitativa |

|vitamin-b1\_100g |DoubleType() |Cuantitativa |

|vitamin-b2\_100g |DoubleType() |Cuantitativa |

|vitamin-pp\_100g |DoubleType() |Cuantitativa |

|vitamin-b6\_100g |DoubleType() |Cuantitativa |

|vitamin-b9\_100g |DoubleType() |Cuantitativa |

|folates\_100g |DoubleType() |Cuantitativa |

|vitamin-b12\_100g |DoubleType() |Cuantitativa |

|biotin\_100g |DoubleType() |Cuantitativa |

|pantothenic-acid\_100g |DoubleType() |Cuantitativa |

|silica\_100g |DoubleType() |Cuantitativa |

|bicarbonate\_100g |DoubleType() |Cuantitativa |

|potassium\_100g |DoubleType() |Cuantitativa |

|chloride\_100g |DoubleType() |Cuantitativa |

|calcium\_100g |DoubleType() |Cuantitativa |

|phosphorus\_100g |DoubleType() |Cuantitativa |

|iron\_100g |DoubleType() |Cuantitativa |

|magnesium\_100g |DoubleType() |Cuantitativa |

|zinc\_100g |DoubleType() |Cuantitativa |

|copper\_100g |DoubleType() |Cuantitativa |

|manganese\_100g |DoubleType() |Cuantitativa |

|fluoride\_100g |DoubleType() |Cuantitativa |

|selenium\_100g |DoubleType() |Cuantitativa |

|chromium\_100g |DoubleType() |Cuantitativa |

|molybdenum\_100g |DoubleType() |Cuantitativa |

|iodine\_100g |DoubleType() |Cuantitativa |

|caffeine\_100g |DoubleType() |Cuantitativa |

|taurine\_100g |DoubleType() |Cuantitativa |

|ph\_100g |DoubleType() |Cuantitativa |

|fruits-vegetables-nuts\_100g |DoubleType() |Cuantitativa |

|fruits-vegetables-nuts-dried\_100g |DoubleType() |Cuantitativa |

|fruits-vegetables-nuts-estimate\_100g |DoubleType() |Cuantitativa |

|fruits-vegetables-nuts-estimate-from-ingredients\_100g|DoubleType() |Cuantitativa |

|collagen-meat-protein-ratio\_100g |DoubleType() |Cuantitativa |

|cocoa\_100g |DoubleType() |Cuantitativa |

|chlorophyl\_100g |DoubleType() |Cuantitativa |

|carbon-footprint\_100g |DoubleType() |Cuantitativa |

|carbon-footprint-from-meat-or-fish\_100g |DoubleType() |Cuantitativa |

|nutrition-score-fr\_100g |IntegerType() |Cuantitativa |

|nutrition-score-uk\_100g |IntegerType() |Cuantitativa |

|glycemic-index\_100g |DoubleType() |Cuantitativa |

|water-hardness\_100g |DoubleType() |Cuantitativa |

|choline\_100g |DoubleType() |Cuantitativa |

|phylloquinone\_100g |DoubleType() |Cuantitativa |

|beta-glucan\_100g |DoubleType() |Cuantitativa |

|inositol\_100g |DoubleType() |Cuantitativa |

|carnitine\_100g |DoubleType() |Cuantitativa |

+-----------------------------------------------------+---------------+----------------+

Luego realizamos el filtrado de la información para obtener 3 tipos de productos que definimos como iniciales, tales como, Seitán, Soja Texturizada y Tofu.

Texto

Descripción generada automáticamente

Luego dimos paso a la limpieza de los datos, en donde:

* Se convienen los datos contenidos en los dataframe a mayúsculas, esto con el fin de estandarizarlos.
* Se eliminan las filas en las que todos los valores son nulos, ya que se trata de filas que no contribuyen en los análisis que se harán en las etapas siguientes.
* Se identificar las columnas que son de tipo string y las columnas que son numéricas. Y se reemplazan los casos en que existe un NaN en cada uno de los dataframe creados por un dato. ("" cuando es String y 0 cuando es Numérica).
* Se eliminan los datos duplicados, si los hubiera, en el dataframe.
* Con el fin de no tener problemas al momento de invocar las columnas, se estandarizan los nombres que actualmente existen en el dataframe.

Imagen que contiene Texto

Descripción generada automáticamente

Luego de este proceso, obtenemos 1206 productos.

Tabla

Descripción generada automáticamente

1. https://world.openfoodfacts.org/data [↑](#footnote-ref-1)
2. <https://opendatacommons.org/licenses/odbl/1-0/> [↑](#footnote-ref-2)
3. https://creativecommons.org/licenses/by-sa/3.0/deed.en [↑](#footnote-ref-3)