Food Facts

Rodolfo Viana

10-02-2016

Utilizando os dados do Open Food Facts database (http://world.openfoodfacts.org/data), um gratuito, aberto e colaborativo database sobre comida ao redor do mundo, foram feitas análises sobre a concetração de cafeina, proteína e sódio nos alimentos.

Nos dados iniciais temos uma tabela que contém as seguintes colunas:

- code (text)
- url (text)
- creator (text)
- created_t (text)
- created_datetime (text)
- last_modified_t (text)
- last_modified_datetime (text)
- product_name (text)
- generic_name (text)
- quantity (text)
- packaging (text)
- packaging_tags (text)
- brands (text)
- brands_tags (text)
- categories (text)
- categories_tags (text)
- categories_en (text)
- origins (text)
- origins_tags (text)
- manufacturing_places (text)
- manufacturing_places_tags (text)
- labels (text)
- labels_tags (text)
- labels_en (text)
- emb_codes (text)
- emb_codes_tags (text)
- first_packaging_code_geo (text)
- cities (text)
- cities_tags (text)
- purchase_places (text)
- stores (text)
- countries (text)
- countries_tags (text)
- countries_en (text)
- ingredients_text (text)
- allergens (text)
- allergens_en (text)

- traces (text)
- traces_tags (text)
- traces_en (text)
- serving_size (text)
- no_nutriments (numeric)
- additives_n (numeric)
- additives (text)
- additives_tags (text)
- additives_en (text)
- ingredients_from_palm_oil_n (numeric)
- ingredients_from_palm_oil (numeric)
- ingredients_from_palm_oil_tags (text)
- ingredients_that_may_be_from_palm_oil_n (numeric)
- ingredients_that_may_be_from_palm_oil (numeric)
- ingredients_that_may_be_from_palm_oil_tags (text)
- nutrition_grade_uk (numeric)
- nutrition_grade_fr (text)
- pnns_groups_1 (text)
- pnns_groups_2 (text)
- states (text)
- states_tags (text)
- states_en (text)
- main_category (text)
- main_category_en (text)
- image_url (text)
- image_small_url (text)
- energy_100g (numeric)
- energy_from_fat_100g (numeric)
- fat_100g (numeric)
- saturated_fat_100g (numeric)
- butyric_acid_100g (numeric)
- caproic_acid_100g (numeric)
- caprylic_acid_100g (numeric)
- capric_acid_100g (numeric)
- lauric_acid_100g (numeric)
- myristic_acid_100g (numeric)
- palmitic_acid_100g (numeric)
- stearic_acid_100g (numeric)
- arachidic_acid_100g (numeric)
- behenic_acid_100g (numeric)
- lignoceric_acid_100g (numeric)
- cerotic_acid_100g (numeric)
- montanic_acid_100g (numeric)
- melissic_acid_100g (numeric)
- monounsaturated_fat_100g (numeric)
- polyunsaturated_fat_100g (numeric)
- omega_3_fat_100g (numeric)
- alpha_linolenic_acid_100g (numeric)
- eicosapentaenoic_acid_100g (numeric)

- docosahexaenoic_acid_100g (numeric)
- omega_6_fat_100g (numeric)
- linoleic_acid_100g (numeric)
- arachidonic_acid_100g (numeric)
- gamma_linolenic_acid_100g (numeric)
- dihomo_gamma_linolenic_acid_100g (numeric)
- omega_9_fat_100g (numeric)
- oleic_acid_100g (numeric)
- elaidic_acid_100g (numeric)
- gondoic_acid_100g (numeric)
- mead_acid_100g (numeric)
- erucic_acid_100g (numeric)
- nervonic_acid_100g (numeric)
- trans_fat_100g (numeric)
- cholesterol_100g (numeric)
- carbohydrates_100g (numeric)
- sugars_100g (numeric)
- sucrose_100g (numeric)
- glucose_100g (numeric)
- fructose_100g (numeric)
- lactose_100g (numeric)
- maltose_100g (numeric)
- maltodextrins_100g (numeric)
- starch_100g (numeric)
- polyols_100g (numeric)
- fiber_100g (numeric)
- proteins_100g (numeric)
- casein_100g (numeric)
- serum_proteins_100g (numeric)
- nucleotides_100g (numeric)
- salt_100g (numeric)
- sodium_100g (numeric)
- alcohol_100g (numeric)
- vitamin_a_100g (numeric)
- beta_carotene_100g (numeric)
- vitamin_d_100g (numeric)
- vitamin_e_100g (numeric)
- vitamin_k_100g (numeric)
- vitamin_c_100g (numeric)
- vitamin_b1_100g (numeric)
- vitamin_b2_100g (numeric)
- vitamin_pp_100g (numeric)
- vitamin_b6_100g (numeric)
- vitamin_b9_100g (numeric)
- vitamin_b12_100g (numeric)
- biotin_100g (numeric)
- pantothenic_acid_100g (numeric)
- silica_100g (numeric)
- bicarbonate_100g (numeric)

- potassium_100g (numeric)
- chloride_100g (numeric)
- calcium_100g (numeric)
- phosphorus_100g (numeric)
- iron_100g (numeric)
- magnesium_100g (numeric)
- zinc_100g (numeric)
- copper_100g (numeric)
- manganese_100g (numeric)
- fluoride_100g (numeric)
- selenium_100g (numeric)
- chromium_100g (numeric)
- molybdenum_100g (numeric)
- iodine_100g (numeric)
- caffeine_100g (numeric)
- taurine_100g (numeric)
- ph_100g (numeric)
- fruits_vegetables_nuts_100g (numeric)
- collagen_meat_protein_ratio_100g (numeric)
- cocoa_100g (numeric)
- chlorophyl_100g (numeric)
- carbon_footprint_100g (numeric)
- nutrition_score_fr_100g (numeric)
- nutrition_score_uk_100g (numeric)

Por se tratar de mais de 50M de dados, foi decidido que iriamos trabalhar apenas com as seguintes colunas:

- product_name
- generic_name
- quantity
- brands_tags
- categories_en
- origins_tags
- stores
- countries_en
- serving_size
- no_nutriments
- additives_n
- main_category_en
- trans_fat_100g
- fiber_100g
- proteins_100g
- salt_100g
- sodium_100g
- calcium_100g
- caffeine_100g
- image_url
- ingredients_text

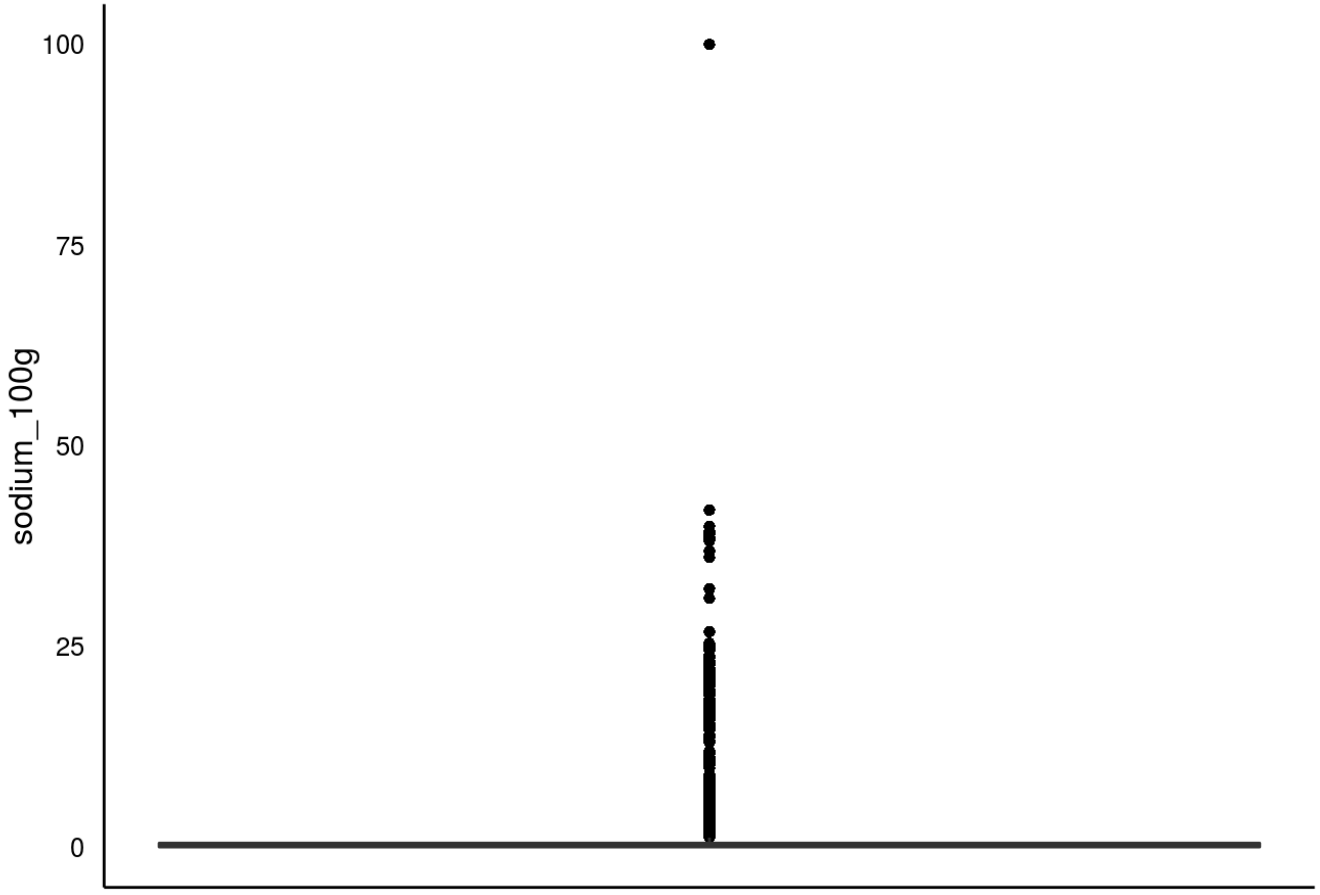
```
library(dplyr)
library(ggplot2)
```

```
file <- read.csv("food_facts.csv", sep=";")

file$sodium_100g <- as.character(file$sodium_100g)
file$sodium_100g[is.na(file$sodium_100g)] <- 0
file$sodium_100g <- gsub(",", ".", file$sodium_100g)
file$sodium_100g <- as.numeric(file$sodium_100g)

file$caffeine_100g <- as.character(file$caffeine_100g)
file$caffeine_100g[is.na(file$caffeine_100g)] <- 0
file$caffeine_100g <- gsub(",", ".", file$caffeine_100g)
file$caffeine_100g <- as.numeric(file$caffeine_100g)

file$proteins_100g <- as.character(file$proteins_100g)
file$proteins_100g[is.na(file$proteins_100g)] <- 0
file$proteins_100g <- gsub(",", ".", file$proteins_100g)
file$proteins_100g <- as.numeric(file$proteins_100g)

file_sodium <- filter(file, sodium_100g != 0)
file_caffeine <- filter(file, caffeine_100g != 0)
file_proteins <- filter(file, proteins_100g != 0)
```

A nossa primeira curiosidade foi descobrir como que era a distribuição de Sódio, Proteína e Cafeina em todos os alimentos. Para uma melhor visualização da distribuição dos dados utilizamos os boxplot. Por se tratar de um banco de dados colaborativo, existe um grande número de valores em branco. Por causa disso, filtramos e retiramos os alimentos com valores NA para proteína, cafeina e sódio.

Para o sódio temos o seguinte boxplot:

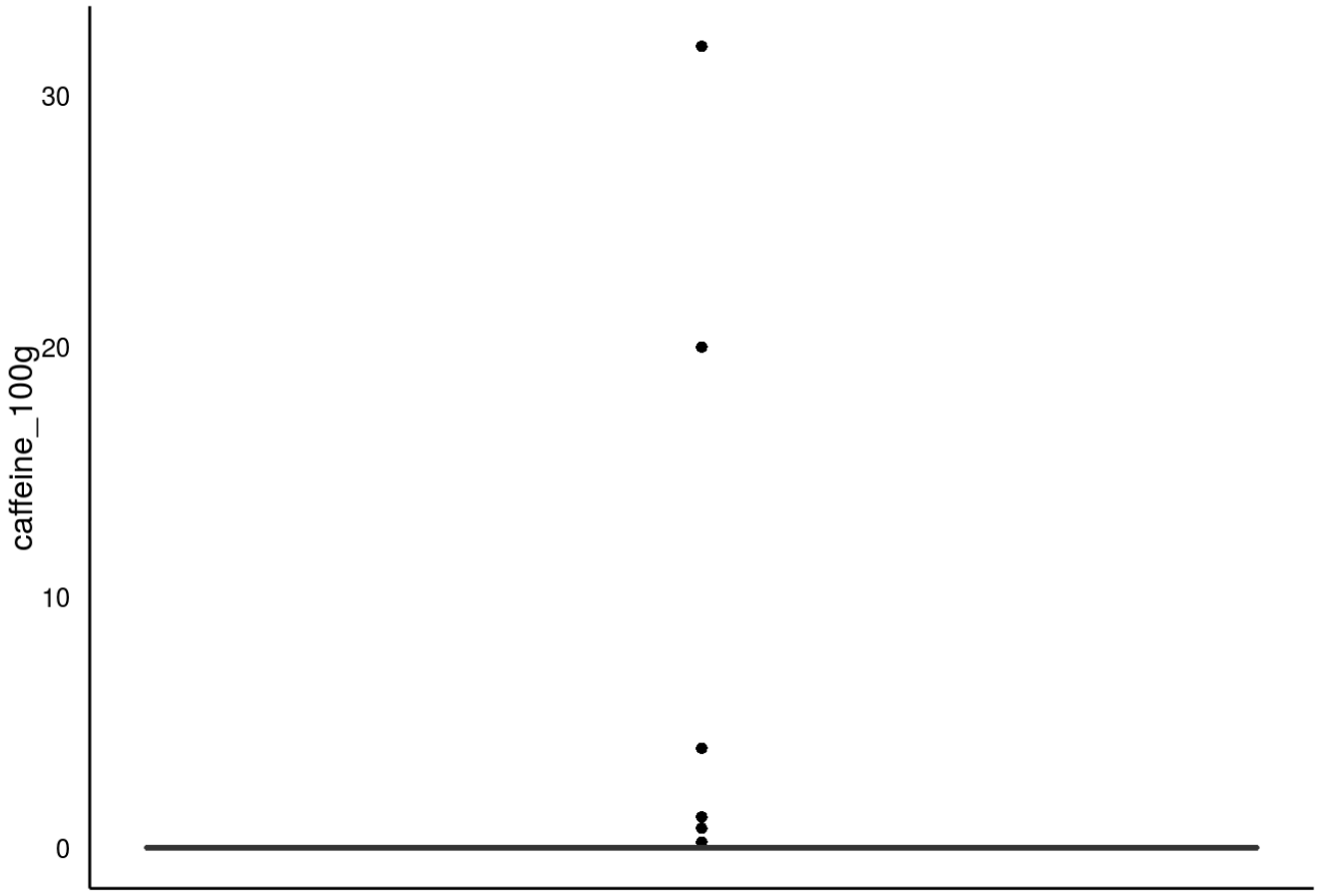
```
ggplot(file_sodium, aes(1, sodium_100g)) +
  geom_boxplot() +
  theme_classic() +
  theme(axis.ticks = element_blank(),
         axis.text.x=element_blank(),
        axis.title.x=element_blank(),
        legend.position="none")
```



É possível notar que existe um grande número de outliers e que tanto a média, 1 e 3 quartil estão próximos a zero. Interessante observar que existe um alimento com mais da metade da concentração de sódio dos demais.

Para a cafeina temos o seguinte boxplot:

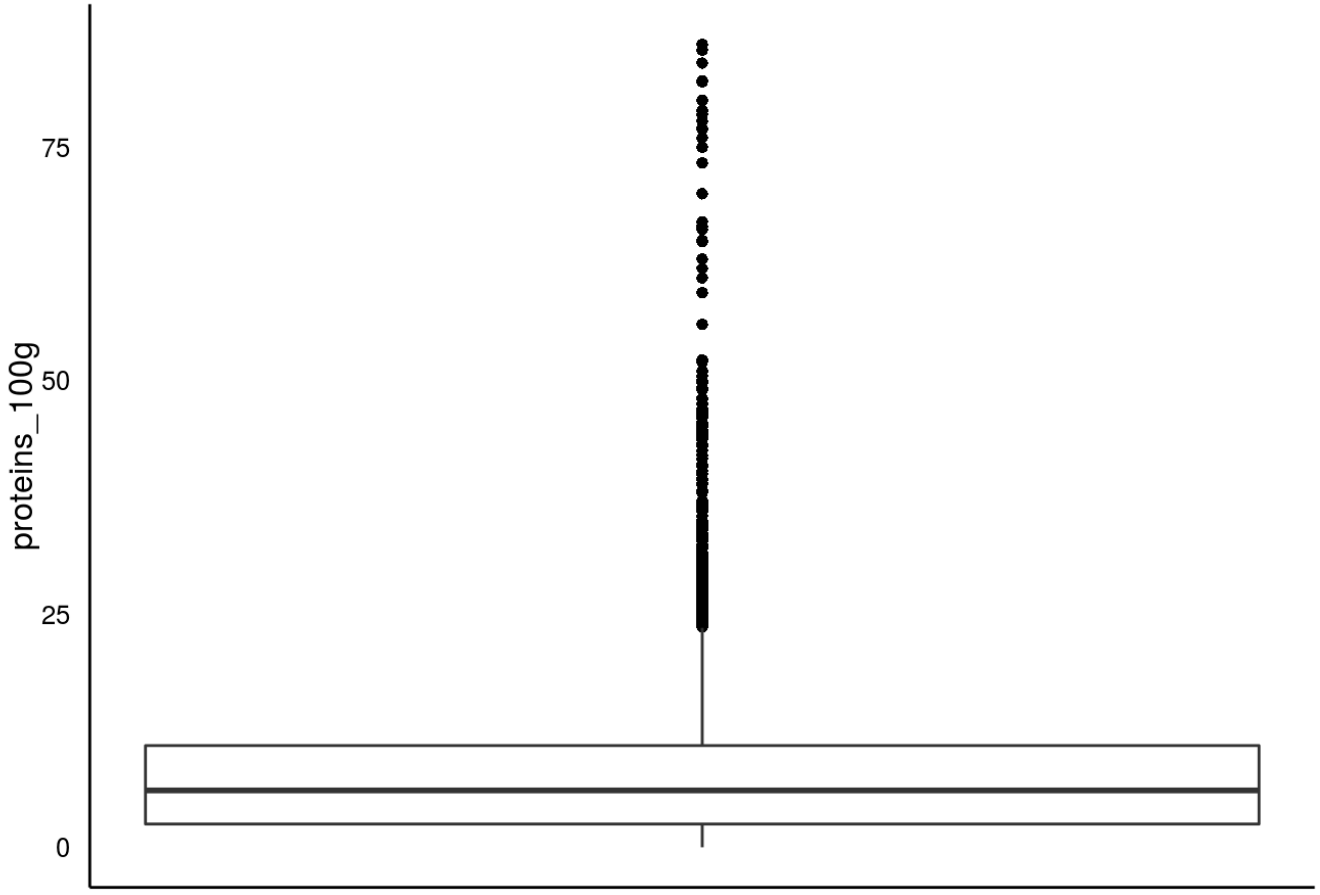
```
ggplot(file_caffeine, aes(1, caffeine_100g)) +
  geom_boxplot() +
  theme_classic() +
  theme(axis.ticks = element_blank(),
        axis.text.x = element_blank(),
       axis.title.x = element_blank())
```



A cafeina possui menos outliers que o sódio, porém a média, 1 e 3 quartil também se encontra perto de zero.

Para a proteína temos o segunite boxplot:

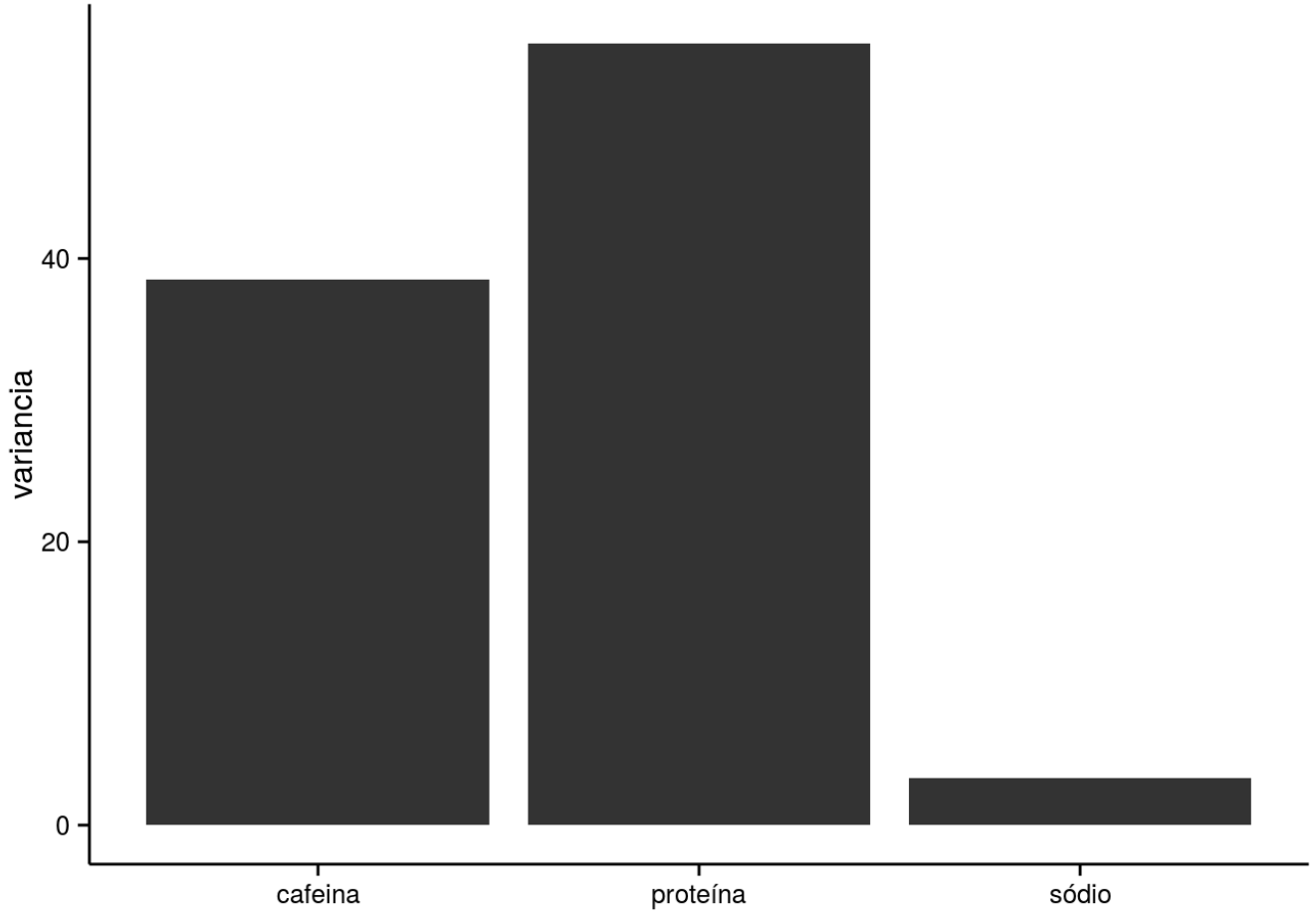
```
ggplot(file_proteins, aes(1, proteins_100g)) +
  geom_boxplot() +
  theme_classic() +
  theme(axis.ticks = element_blank(),
        axis.text.x=element_blank(),
       axis.title.x=element_blank(),
       legend.position="none")
```



A proteína, diferente das outras substâncias, possui um boxplot com 1 e 3 quartil com valores diferentes. Mostrando dessa forma que a proteína possui diversos alimentos com diferentes grau de concentração

A nossa segunda curiosidade foi descobrir qual das três substancias possui maior variância nos alimentos

```
df <- data.frame(tipo = c("sódio", "cafeina", "proteína"),
                 variancia = c(var(file_sodium$sodium_100g), var(file_caffeine
$caffeine_100g), var(file_proteins$proteins_100g)))

ggplot(df, aes(x = tipo, y = variancia)) +
  geom_bar(stat = "identity") +
  theme_classic() +
  theme(axis.title.x = element_blank())
```



É possível notar que a proteína é a substância que possui maior variância e o sódio menor variância de concentração nos alimentos

A nossa última curiosidade foi descobrir qual alimento possui a maior concentração de Sódio, Proteína e Cafeina

Para a cafeina temos o alimento:

```
file_caffeine <- file_caffeine[order(-file_caffeine$caffeine_100g),]
file_caffeine[1,]$product_name
```

```
## [1] Red Bull energy drink
## 50932 Levels:    알 \U0001f37a 통깨 짜왕 ゆず 미역 울무차 콩두유 ... 黑瓶眼药水
```

foto do alimento (http://en.openfoodfacts.org/images/products/22220768/front.8.400.jpg)

O que não foi nenhuma surpresa, já que o RedBull é conhecido por ser um alimento que pode te proporcionar uma “energia extra”

Para a proteína temos:

```
file_proteins <- file_proteins[order(-file_proteins$proteins_100g),]
file_proteins[1,]$product_name
```

```
## [1] Blattgelatine weiss
## 50932 Levels:    알 \U0001f37a 통깨 짜왕 ゆず 미역 울무차 콩두유 ... 黑瓶眼药水
```

```
file_proteins[1,]$image_url
```

```
## [1] http://en.openfoodfacts.org/images/products/20153465/front.8.400.jpg
## 61200 Levels:  ...
```

foto do alimento (http://en.openfoodfacts.org/images/products/20153465/front.8.400.jpg)

Uma especie de gelatina vendida na Alemanha

Para o sódio tivemos uma surpresa, pois o alimento cadastrado como tendo a maior concentração de sódio não possuía praticamente nenhum registro (nome, localidade, foto, etc). Por se tratar de um conjunto de dados aberto e colaborativo isso é considerado “normal”. Por essa razão resolvemos observar o alimento na segunda posição:

```
file_sodium <- file_sodium[order(-file_sodium$sodium_100g),]
file_sodium[2,]$product_name
```

```
## [1] Himalayan Pink Salt
## 50932 Levels:    알 \U0001f37a 통깨 짜왕 ゆず 미역 울무차 콩두유 ... 黑瓶眼药水
```

Foto do alimento (http://en.openfoodfacts.org/images/products/009/661/991/1936/front.6.400.jpg)

O alimento encontrado foi um sal vendido nos Estados Unidos