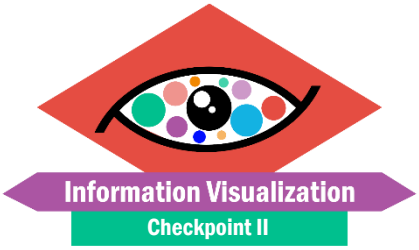

Checkpoint II: Data Cleaning & Processing

Group: G09
Date: 2022/10/04

Initial Dataset

The initial dataset had all the nutritional values for **around 8.8k items** (food products). It was a **table** with a total of **77 columns**, most of them containing a single value and its unit of measurement, which represented the quantity of each component. An example of this dataset is as follows:

```
3,"Teff, uncooked",Teff,uncooked,Vegetables,100 g,367,2.4g,0.4g,0,12.00mg,13.1 mg,
0,0, 3.363 mg,0.942 mg,0.270 mg,0.390 mg,9.00 IU,0.00 mcg,0.00 mcg,5.00 mcg,0.00
mcg,66.00 mcg,0,0,0.482 mg,0,0,0.08 mg,0.08 mg,1.9 mcg,180.00 mg,0.810 mg,7.63
mg,184.00 mg,9.240 mg,429.00 mg,427.00 mg,4.4 mcg,3.63 mg,13.30 g,0.747 g,0.517
g,0.820 g,0.236 g,3.349 g,0.477 g,0.301 g,0,0.501 g,1.068 g,0.376 g,0.428 g,0.698
g,0.664 g,0.622 g,0.510 g,0.139 g,0.458 g,0.686 g,73.13 g,8.0 g,1.84 g,0.47 g,0.00
g,0.73 g,0.00 g,0.01 g,0.62 g,2.38 g,0.449 g,0.589 g,1.071 g,0,0,2.37 g,0,0,8.82 g
```

Selected/Derived Data

From the initial dataset we decided to consider only the most known and relevant columns as well as those we decided to add due to their relevance. We **maintained columns** such as *name, type, description, category, serving size, calories, protein, saturated fat, vitamin-b12, carbohydrates*, etc.

For **derived measures**, we decided to compute *fat_percentage*, *water_percentage_serving_size* and *protein_percentage* for each item. These measures are useful to compare, contrast and correlate different types of products. We also added measures comparing the amount of a certain attribute with the recommended daily intake of said attribute. These added measures were created for *calories, protein, sodium* and *carbohydrates* adding the suffix _daily_intake_percentage. It is important to note that some values of the dataset (e.g. *calories*) are **already derived data** (can be calculated taking in consideration the grams of protein, fat and carbohydrates).

Data Abstraction

Our dataset is organized in **tables**. The *name, type, description* and *category* are **nominal** attributes. Hence the remaining values are all quantities (e.g. *serving_size, protein, vitamin_c, sodium*, etc.) and percentages (e.g. *fat_percentage*, *protein_daily_intake_percentage* and the other derived measures), they are all **ratio** variables. Each one of them has a **true zero** and any two values have a **meaningful ratio**, making the operations of multiplication and division meaningful. All of these variables have a **sequential scale** since the numeric values increase continuously, starting at **zero**.

Data Processing

Initially, we had 8.8k items in our dataset and some had **missing values**. To tackle some of them, we considered **mathematically computing** the missing values but since it was a negligible percentage of our dataset (around 6%) we decided to just **drop** them. For **outliers**, we decided to keep every item since there are in fact items who have extremely high quantities of certain attributes (e.g. oils are composed essentially of fat, some of them around 99%). Afterwards, we decided to remove manually some columns that were not as relevant as the others. By doing that, we noticed that there were many food items with nearly the same composition and name. Others were very specific and negligible towards the goal of this project, so we decided to **remove** them. All of this has been done using the **Pandas library** in Python.

We also noted that there were few ways of grouping items of the same category and comparing them as a group. Therefore we decided to categorize each item **manually** by adding a **new column** with their *category* (e.g. Vegetables, Meat, Dairy, etc.). Regarding the **derived measures**, various columns were added after **cross-referencing** with the reference table and calculating the percentages of each attribute per item.

Mapping (Data sample/Questions)

```
,name,type,description,category,serving_size,calories,calories_daily_intake_percent
age,total_fat,fat_percentage,saturated_fat,cholesterol,sodium,sodium_daily_intake_p
ercentage,vitamin_b12,vitamin_b6,vitamin_c,vitamin_d,calcium,magnesium,potassium,pr
otein,protein_percentage,protein_daily_intake_percentage,carbohydrate,carbohydrate_
daily_intake_percentage,fiber,sugars,fructose,glucose,lactose,fat,alcohol,caffeine,
water,water_percentage_serving_size
```

```
805,"Ground turkey, raw, fat free",Ground turkey,"raw, fat free",Meat,100 g,112, 5.6%,
2g, 16.1%,0.5g, 55mg,51.00 mg,1%,0.51 mcg,0.857 mg,0.0 mg,14.00 IU,3.00 mg,29.00
mg,295.00 mg,23.57 g,84.2%, 27.8%,0.00 g,0%,0.0 g,0.00 g,0,0,0,1.95 g,0.0 g,0.00
mg,74.66 g,74.7%
```

```
(from "reference.csv")
,tot_calories,protein(g),protein(kcal),carbohydrates(g),carbohydrates(kcal),fat(g),
tot_fat(kcal),sat_fat(kcal),sugar(kcal),potassium(g),sodium(g)
```

```
Adult,2000,84.86,300,340,1300,62.9,600,200,200,0.3,5
```

- Do food items with a high **protein percentage** and **low calories per serving** generally have a **low percentage of saturated fat**?

We can correlate the *protein_percentage, calories* and *serving_size* with the *fat_percentage*, more specifically with the quantity of *saturated_fat* for every item in the desired conditions. Given the first example, check if the item has a high protein percentage (84.2%) and low calories per serving (112 kcal per 100g) and verify if it has a low percentage of saturated fat (it is 16.1% fat and has 0.5 grams of saturated fat).

- How does the **total fat** and **sodium** influence the **cholesterol** of an item?

Relate the total fat of a food item (*total_fat*) and sodium (*sodium*) with the cholesterol (*cholesterol*) for an item and search for a direct correlation.

- Is **fish** generally **less caloric** than **meat** with **a low fat percentage**?

Taking the items with the *category* Fish with lower calories and compare their fat percentage with the items with the *category* Meat.

- Does **fiber** rich **starchy food** tend to have less **cholesterol**?

From the items in the Starchy food category, we can filter for those with the most fiber and compare those values with the values in the cholesterol column.

- Is **water** rich food **healthier**?

Between the items with a high *water_percentage_serving_size* (derived measure) we can take into account the macronutrients and sodium, as well as their percentages and compare them with the reference table items.

- Does **caffeine** reduce the presence of **B-vitamins**, **vitamin C** and **minerals** in the food?

Taking into account the items in which caffeine is present, we can explore the quantities of B-vitamins (*vitamin_b6, vitamin_b12*), vitamin C (*vitamin_c*) and minerals (*sodium, potassium, calcium* and *magnesium*).