Nutrition Parameter Analysis

Group 4:
Anton Wohlgemuth (11778996@s.wu.ac.at)
Michał Maternicki (1552836@s.wu.ac.at)
...

Overview

The objective of our project was to use the information we found on the USDA Food Composition Database to answer some question regarding nutrition, that we asked ourselves.

- Are sweets really that much higher in sugars than vegetables?
- Is the amount of kcal correlated to the amount of sugar in certain products?
- How does the fat composition look like in different food groups?

Besides these question we also wanted to check some correlations and furthermore cluster our food products by different parameters, to maybe discover some new type of food group.
In order to better analyze our food data we manually labeled them into food groups.
(Vegetables, Fruits, Meats, Milk Products, Sweets, FastFood, Beverages)
Last but not least we also tried to come up with a classification model that can classify the type of a specific food product based on their nutritional parameters.

Research questions

- Are the following assumptions true?

1. Foods high/(low) in Sugar are also high/(low) in calories
2. Foods high/(low) in Fats are also high/(low) in calories
3. Foods high/(low) in Fibre contain a high amount of Vitamins
4. Foods high/(low) in Fats are high/(low) in Cholesterol
5. Foods with high water content are low in calories

- How does the fat composition looks like in different food groups?
- Are sweets really that high in sugars?
- Can we cluster the food products by certain parameters?
- Can we create a model, which can classify a food product based on its nutrition parameters? (Kcal, Carbohydrates, Protein, Fats)

Dataset(s)

The dataset which is being used is taken from the USDA Food Composition Database. It contains the nutrition components (e.g., water, sugar, calcium, etc) of 8790 types of food. The dataset can be downloaded here: (https://ndb.nal.usda.gov/ndb/). The dataset was delivered without any filtering or cleaning.

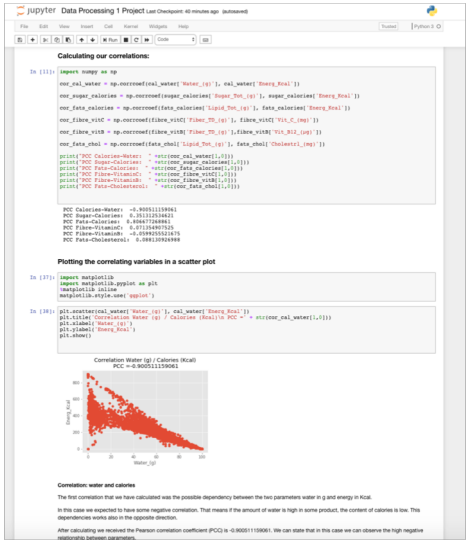
Steps

1. Reading the data
2. Check if the dataset is tidy
3. Inspecting the data (which structure, formats, etc.)
4. Checking and handling duplicates
5. Checking and handling missing data
6. Calculating correlations
7. Visualizing correlations
8. Clustering the data (+ visualization)
9. Classification based on nutrition parameters.

Tools

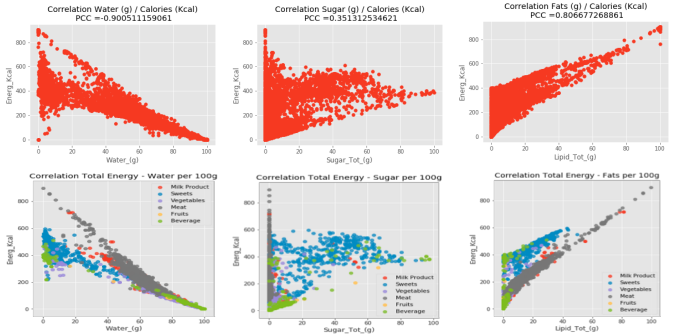
python, plotly, jupyter, Seaborn, matplotlib, scikit learn, pandas

Structure of the notebook

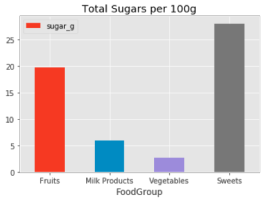


Results:

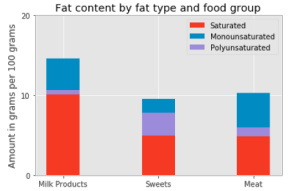
Correlations:



Where can we find the highest amount of sugar per 100g?



What types of fats do we find in different food groups?



Limitations and Lessons Learned

One of our goals was to cluster our data by some nutritional parameters, the most interesting thing was that the statistical methods we used suggested us that there are 3-4 different clusters, but after we visualized them we saw that these cluster didn't make any sense. This was really interesting to see.
After we manually labelled our data into food groups and analyzed the data based on their food group, we could find some interesting insights. For example, we all know that sweets are really high in sugars but so are fruits.
We were also really happy with to visualize different food groups in our correlation plots, since this added some great details to our plots.

Future work:

Merge our dataset with others that contain nutrition parameters of different allergens.
We could then use this information to give recommendation for food alternatives.
Trying other clustering algorithms
Improving our classification model