Data Stewardship
SS 2020

Andrea Siposova
01205149

Part 1

Experiment Description

This assignment consisted of three tasks; clustering, finding correlations between attributes of various foods, and comparing various food categories.

Dataset Description

Data used as input for this experiment comes from the U.S. Department of Agriculture and contains nutrient values of several foods. It contains more than 8000 different foods, and for each of these foods, the content of various minerals, vitamins and their other nutritional features (caloric value, protein content etc.) The file that was provided for this assignment is attached (*‘USDA_Food_database.csv’*).

Task 1 - Clustering

Magnesium and calcium, when ingested at the same time, might negatively affect absorption of one another by our bodies. For example, for people who require a higher daily intake of magnesium, it is also important to not consume it with large amounts of calcium. Therefore, for the first task, the data is clustered based on these two attributes. The motivation behind it was to distinguish which foods should be for example avoided, which should be eaten together and which foods should be eaten at different times in order to not interfere with the absorption of these minerals.

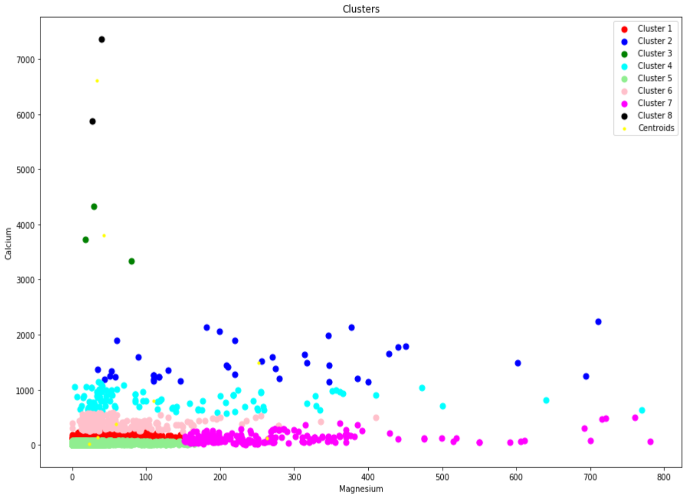

Figure 1

The data was preprocessed and the attribute values were aggregated per category. The mean value of the content of each mineral was used in each category. This dataframe was saved to a csv file (‘mag_cal_per_category.csv’) This clustering was not entirely useful, so clustering was performed on all foods (not aggregated). Shown in *Figure 1*. The results were saved as a csv file (‘*clustered_food.csv’)*. The attributes included are name, category, magnesium and calcium content, and the cluster number of each food).

Task 2 - Correlation

For the second task, preprocessing of the data was very important, since some of the numbers were in the form of strings and used a comma instead of a decimal point. These values had to be corrected and converted, in order to be usable. Otherwise, the correlation analysis would lead to misleading results.

Three interesting findings are that:

- The protein content seems to be positively correlated with the content of monosaturated fatty acids in foods (The more protein, the less monosaturated FA)
- The carbohydrate content is negatively correlated with the water content
- Protein content seems to be negatively correlated with the amount of sugar in food

The produced correlation matrix is attached, as well as the heatmap which was created using seaborn.

Task 3 - Comparison

The three groups chosen for comparison are:

- Vegetables vs Fruits
- Processed foods vs Vegetables
- Animal Fat vs Vegetable Fat

The data was aggregated into categories and descriptive statistics were calculated. This dataframe was saved as a csv file (‘categories_descriptive.csv’). The mean value for each attribute in each category was subsetted.

The pandas class (which is based on matplotlib) was utilized for creating bar charts.

It is important to note that the different bar charts are not proportional to each other since they use different units (e.g. grams and micrograms). It was necessary to keep the units as original, otherwise the bars would not be visible.

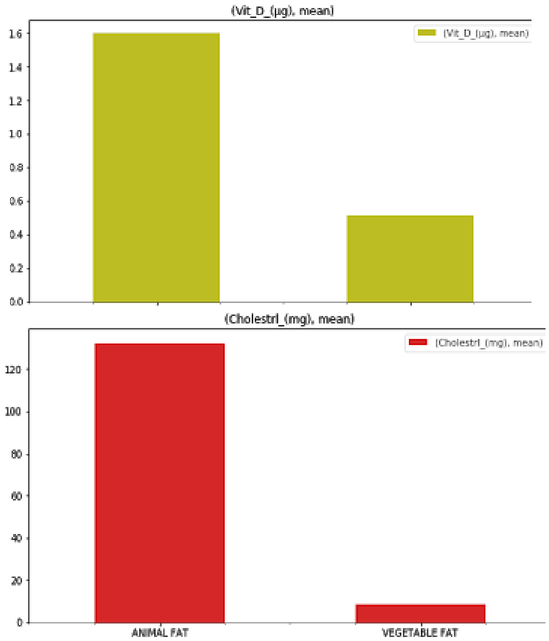

Figure 2

Figure 2 shows 2 examples of how two the categories Animal Fat and Vegetable Fat were compared on an attribute. We can clearly see that animal fat contains about 13-times more cholesterol than vegetable fat. But when it comes to vitamin D, animal fat is the winner - it contains more than 3-times more of it than vegetable fat.

All other visualizations are attached.

Another file shows the data flow and the steps taken in the project to transform our input data into the output files.