Data Analytics **2023**

Names: Erifeoluwa Jamgbadi and Mehrazin Yazdani

Clustering on Food

Due date: 03.06.2023

---

1. Introduction

In this project, we want to preform data analysis on a food dataset. We explore and analyse the data and also draw conclusions from the questions we formed and analysis. Our food dataset has different features representing food facts. We make clusters using the datasets attributes.
Clustering is an unsupervised learning algorithm, which is great for grouping things that have similar things in common. There are many different clustering algorithms that can be used to make "clusters". In our project, we picked only 3 different clustering algorithm to implement, discuss and evaluate, which can be found in more detail in other section in our report.

When we were researching which clustering algorithm to implement, KMeans came up a lot on top clustering algorithms. That's one of the reasons why our initial thought process was that KMeans will be the best clustering algorithm compared to the others we are using, but we can't conclude yet, that is until we see the results from comparing these algorithms.We chose HBDScan as our other clustering algorithm because it is density based and easy to parameterize.We chose this over more famous db scan because it is able to manage clusters with different densities and it doesn't require to select the eplison parameter which is hard to find. The final clustering algorithm that we chose is CURE.This clustering algorithm was mentioned in the course, so we though we would use it.

2. Data exploration

In this section, we describe and explore the data. The following sections shows questions we asked ourselves when exploring the data and answers to those questions.

Question 1 - What does the data look like?

As the data size is too large, only a small representation could be shown of the data. Table 1 shows a few rows of the data and how the dataset looks. There are a total of 78246 rows and 159 columns in the dataset. The columns contain information such as creator, manufacturing places, categories, countries, additives, various ingredients, existing vitamins, etc.

Question 2 - Are there any duplicate rows?

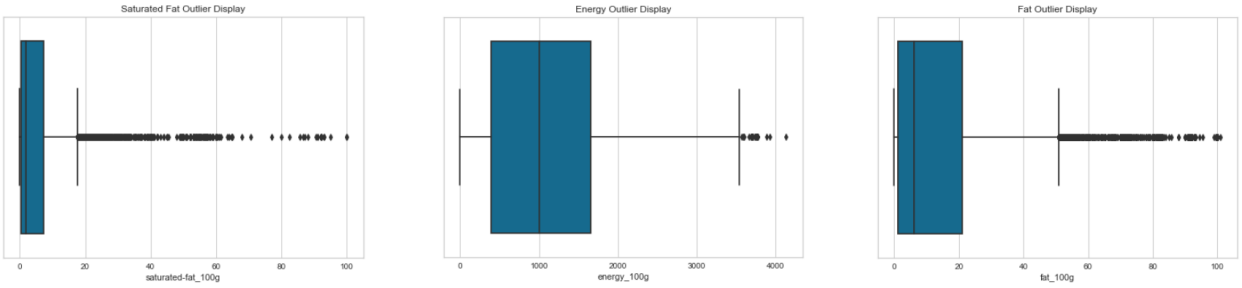
We wanted to check if there were any duplicate rows and we confirmed that there was no duplicate rows.

code	...	creator	created_t	ncreated_datetime	...	nutrition-score-uk_100g
0000000024600	...	date-limite-app	1434530704	2015-06-17T08:45:04Z	...	NaN
0000000027205	...	tacinte	1458238630	2016-03-17T18:17:10Z	...	NaN
0000000036252	...	tacinte	1422221701	2015-01-25T21:35:01Z	...	NaN
0000000039259	...	tacinte	1422221773	2015-01-25T21:36:13Z	...	NaN
0000000039529	...	teolemon	1420147051	2015-01-01T21:17:31Z	...	NaN
0000001071894	...	bcatelin	1409411252	2014-08-30T15:07:32Z	...	NaN
...	...	...	...	...	...	...

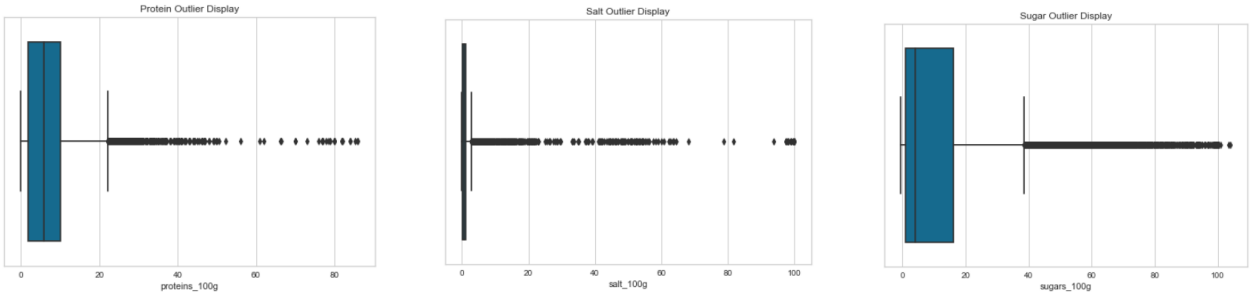
Table 1: Representation of the data

Question 3 - Are there any outliers in the dataset?

Yes, there are outliers the dataset. This result is gotten only after we dropped irrelevant columns and filled in the missing values, which will be discussed in more detail later on in the report. Carbohydrate is the only nutrient that does not have any outliers, so it was left out from Figure 1, there are many outliers in the other nutrients.



(a) Outliers in saturated fat, energy, fat (order of images)



(b) Outliers in protein, salt, sugar(order of images)

Figure 1: Outlier Display of variables

Question 4 - Are there any missing values in the data set?

Yes, we discovered that there were data missing in the dataset. As mentioned before the dataset has 78246 rows and some of the columns in the dataset had 78246 missing data e.g montanic-acid_100g, ingredients_that_may_be_from_palm_oil. This posed as a problem because majority of the dataset

had missing values.

Question 5 - Can we remove those columns with missing values? Are they important?

These are the columns where all their rows have missing data : "nutrition_grade_uk, chlorophyl_100g, erucic-acid_100g, lignoceric-acid_100g, caproic-acid_100g, nervonic-acid_100g, cerotic-acid_100g, melissic-acid_100g, ingredients_from_palm_oil, elaidic-acid_100g, no_nutriments, ingredients_that_may_be_from_palm_oil, mead-acid_100g, butyric-acid_100g, montanic-acid_100g".

Most of the acid have all their row containing missing values. Although, we could decide to remove the acid group and not form a cluster with those columns, when we widened the search, there were still a lot of important columns that have a lot of missing. Therefore, we concluded, that the missing values in the columns can't be ignored.

3. Pre-processing Data

Question 6 - How do we handle and fill unknowns if there are any?

We decided not to use interpolation, mean, median, mode or other similar techniques to fill in the data because the number of missing data was too large and in our case won't be useful for making clusters using the nutritional data. We decided that it would be best to reduce the columns, therefore reducing the amount of missing data rows that would be dropped when dropping rows with missing values. We dropped the columns that had the most missing values and we kept reducing the columns from there. That's how our dataset was reduced from 78246 rows and 159 columns to 36732 rows and 13 columns.

Question 7 - Do the columns/nutrients we picked have any correlation with each other?

Looking at the correlation map in Figure 2, we can see that overall the correlation is not as much as a whole food groups, but there are still some correlation between some of the columns. For this reason, we will look more indepth at these with high correlation as they could answer some questions we have at a later stage.

The highest correlation is 1, but we don't count those ones, as the only ones with 1's are the diagonals(column have largest correlation with themselves, which goes without saying). The next highest correlation is 0.79, which responds to fat and energy with each other. Since, fat and energy have the highest correlation, we can concluded, that the more fat that there is in the food, the higher the energy. Energy seems to have good correlation standing compared to the other nutrients, which is another conclusion that we can make, which is that depending on the nutrients in the food, it will give energy, except for salt.

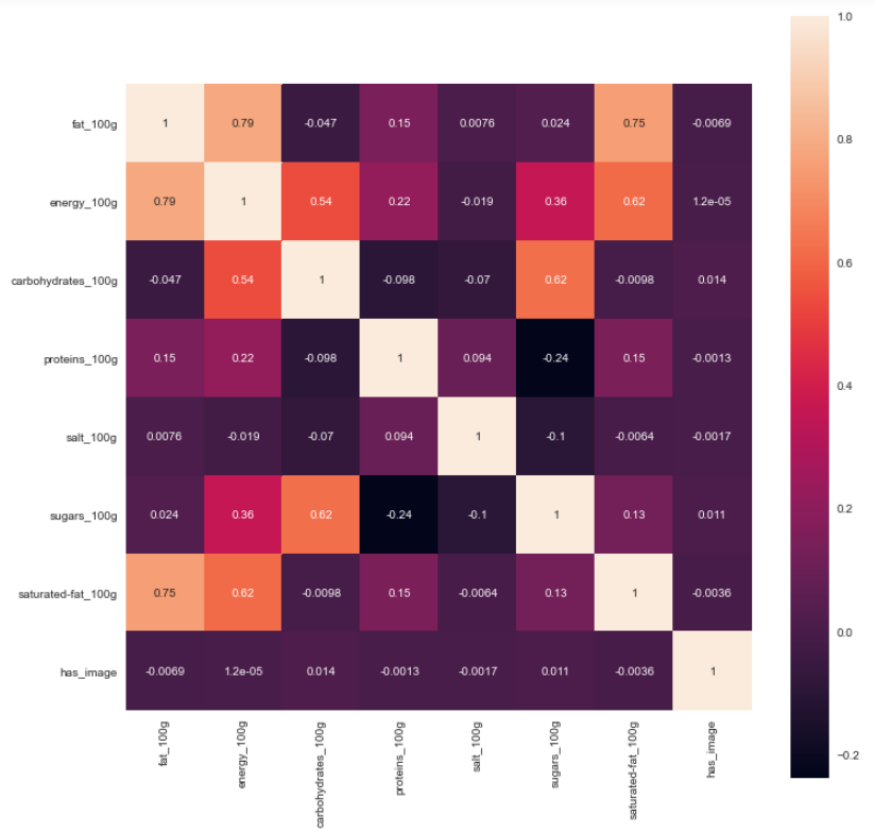


Figure 2: Correlation Map of new dataset

Question 8 - What is the popular category for food?

Plant based food and beverages are the most popular food category according to Figure 3. If France is the most popular country for food, we can conclude that France has healthy foods.

Question 9 - What country has the highest entry of food in the dataset?

As we can see from Figure 4, France is the highest by far, showing that there are more food from France in the dataset by far, either that or had the most complete data in our dataset, making the representation higher.

Question 10 - Which country consumes the most energy for food?

As can be seen in Figure 5, United states and Italy food contains the most energy. We can conclude that their foods contain a lot of fatty nutrients because as can be seen from the correlation map in Figure 2, fat and energy has a high correlation with each other.

Question 11 - Any interesting observations?

Salty foods record high energy as shown in Figure, but on the correlation map, it has really low correlation with each other. We concluded, that it must be there there are other nutrients that have a high correlation with energy that are contained in salty foods such as fat nutrient.

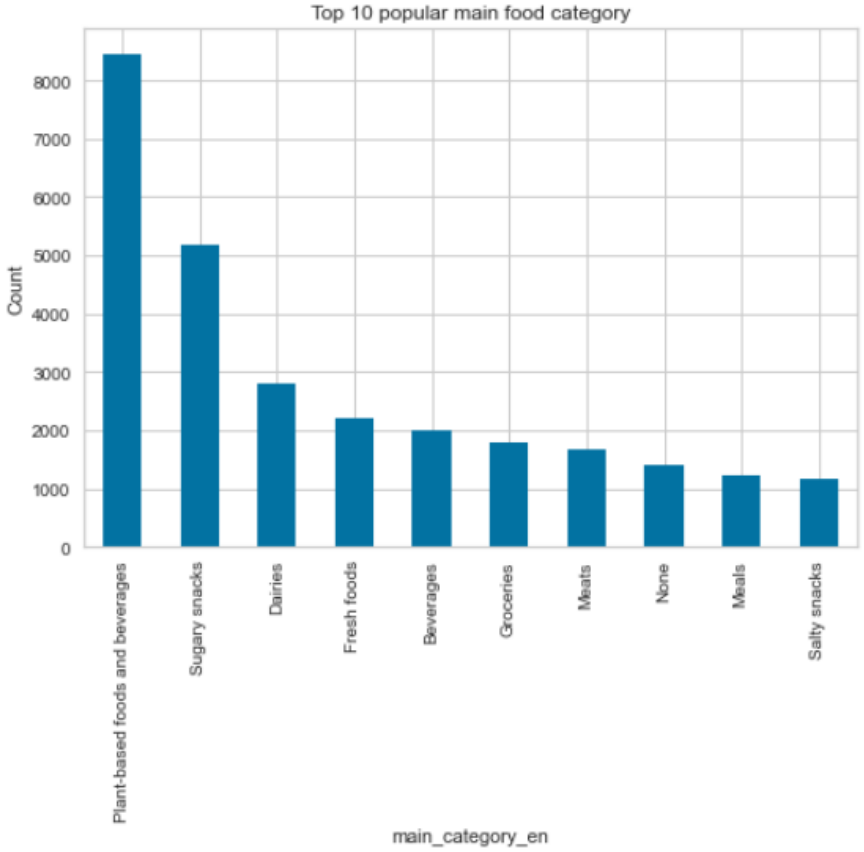


Figure 3: Most popular food in dataset

4. Clustering Algorithms

In this section, we use three clustering algorithms and compare them against each other. The clustering algorithms that we chose are KMeans, HBDScan and CURE.

Question 11 - What is the most optimal number of clusters for the algorithms used?

We found out when using the kmeans clustering algorithm that 6 is the most optimal number of clusters as can be seen from Fig 2. The CURE algorithm had a bad distribution. It clustered almost all data to 2, so we didn't further analyse it, since we already know the performance is bad based on that. HBDScan has an optimal cluster of 5.

Question 12 - Where are the centroids of kmeans located in the cluster distribution?

KMeans algorithm uses centroids in order to assign the points of a cluster based on the nearest centroid. We wanted to see in our data that is being clustered by KMeans, where the location of the centroids where that are used by the kmeans algorithm. Figure 8 shows exactly where the centroids are located for the fat and energy nutrients, so it was able to cluster the points based on the centroid location. There are more examples in the jupyter notebook of the code of different clustering of nutrients and showing the centroid with KMeans.

Question 13 - Does the algorithms cluster the nutrients differently?

Yes, referring to Figure 9, we are able to see how the clustering algorithms clusters the nutrients differently. We can see there are some changes in size for some of the nutrients in each cluster based on the algorithm. In our opinion, the distribution of KMeanns, as well as HDBScan looks more

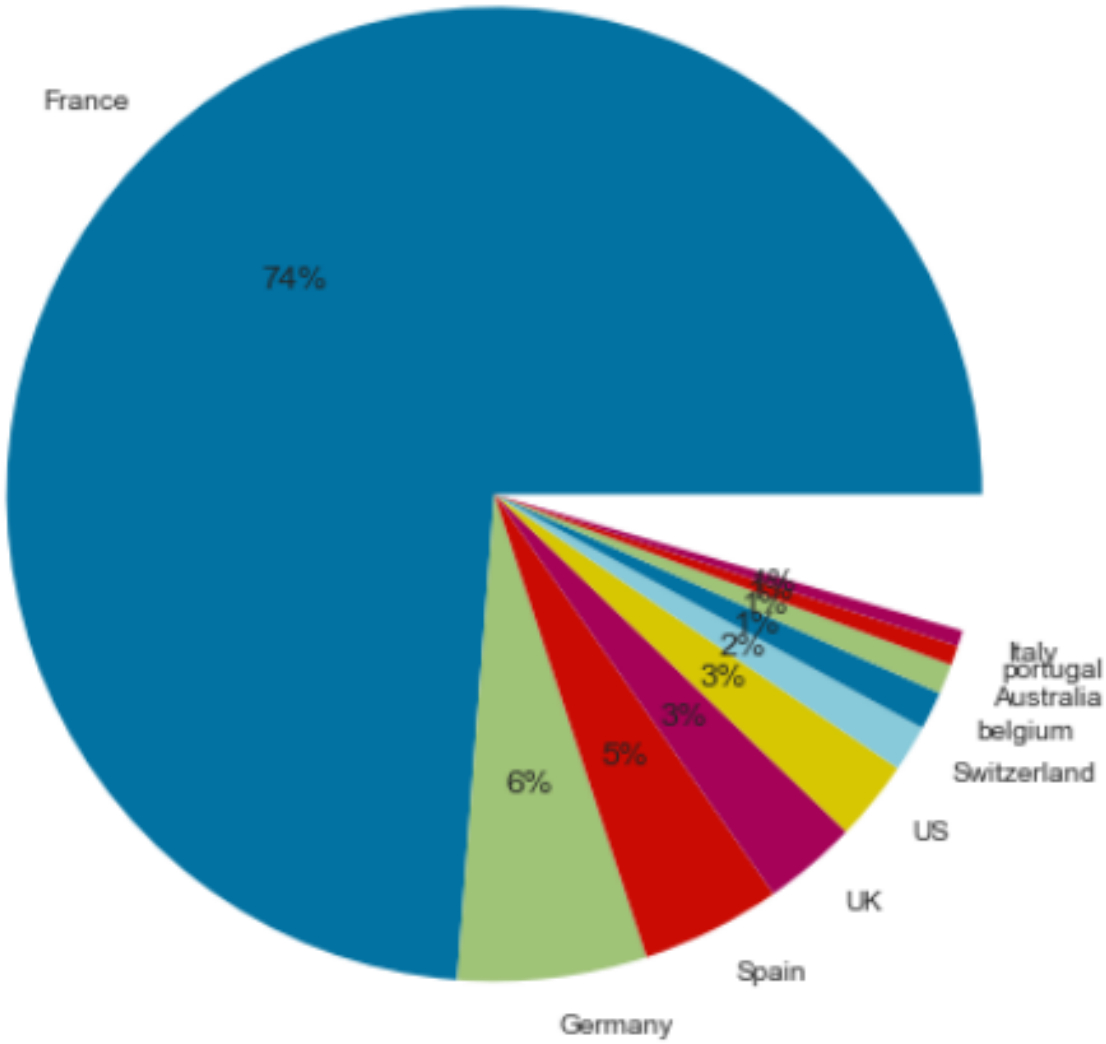


Figure 4: Countries with the highest entries in the dataset

correct. Carbohydrates have a high clustering in both kmeans and hdbscan, but it is no where to be found in CURE. We would say that kmeans and bdbscan are more similar. It seems in the 3rd cluster, all algorithms, recorded low points.

5. Evaluation and Comparison of Model Performance

In this section, we will discuss the evaluation of the model and compare their performances. To evaluate each of the model performance, we did distance calculation.

Question 14 - What is the silhouette score?

The silhouette score tells us about how good the technique is of the clustering algorithm is. The range is between -1 and 1. The worst value is -1, well the best is 1. For performance evaluation, we got the silhouette score for each of the algorithm. The score is as following: KMeans silhouette score is 0.41366574102719933, HBDScan is 0.09455408843161352 and CURE is 0.5430084113676102. CURE has a higher silhouette score than KMeans even though the clustering for CURE was bad. Our opinion on why this could have occurred is because CURE is good at handling outliers in data, so the outliers in the data might have made the score of the CURE clustering algorithm higher than KMeans.

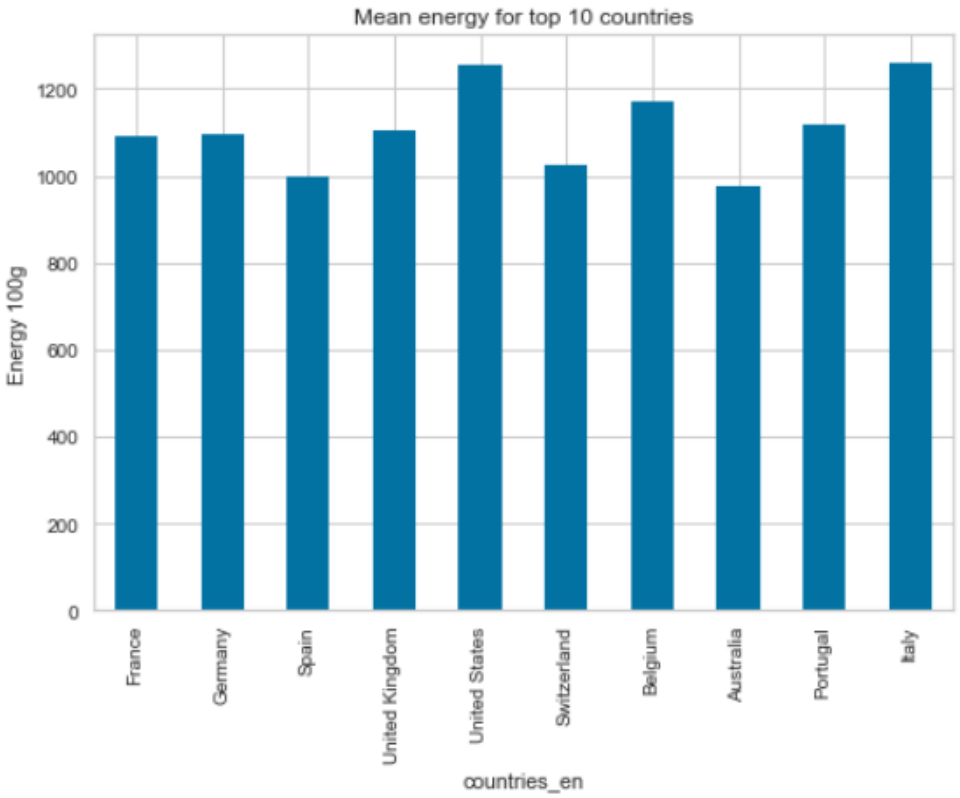


Figure 5: Countries whos food contains the most energy

Question 15 - Which model has the best performance?

We can say from the results that the model with the best performance is KMeans. It was faster than the rest when clustering and it produced a good clustering distribution based on its centroid. CURE algorithm is really slow. Clusters with CURE have a bad distribution, with almost all data in cluster 2, so we didn't further analyse on this algorithm, as we saw the distribution was bad. When running the code, we noticed that it took really long for CURE to run, while kmeans was fast.

Further analysis was done on kMeans as it produced good silhouette results. We preformed normalisation on the kmeans data because we want the data to be all on the same scale because looking at the Figure 9, we realised the calories would be in an example measurement of 1000 and sugar 20grams, which may make the model think that one nutrient is more important than the other.

Question 15 - Summerization table of results

Here is a summerization table of clustering the algorithm.

Clustering Algorithm	Time Complexity	Running Time	Silhouette Score	Our Rating(Best, Medium, Worst)
KMeans	O(NTK)	0.43	0.4136657410271993	Best
HBDScan	O(nlog n)	6.45	0.0945540884316135	Medium
CURE	O(n2 log n)	not taken	0.5430084113676102	Worst

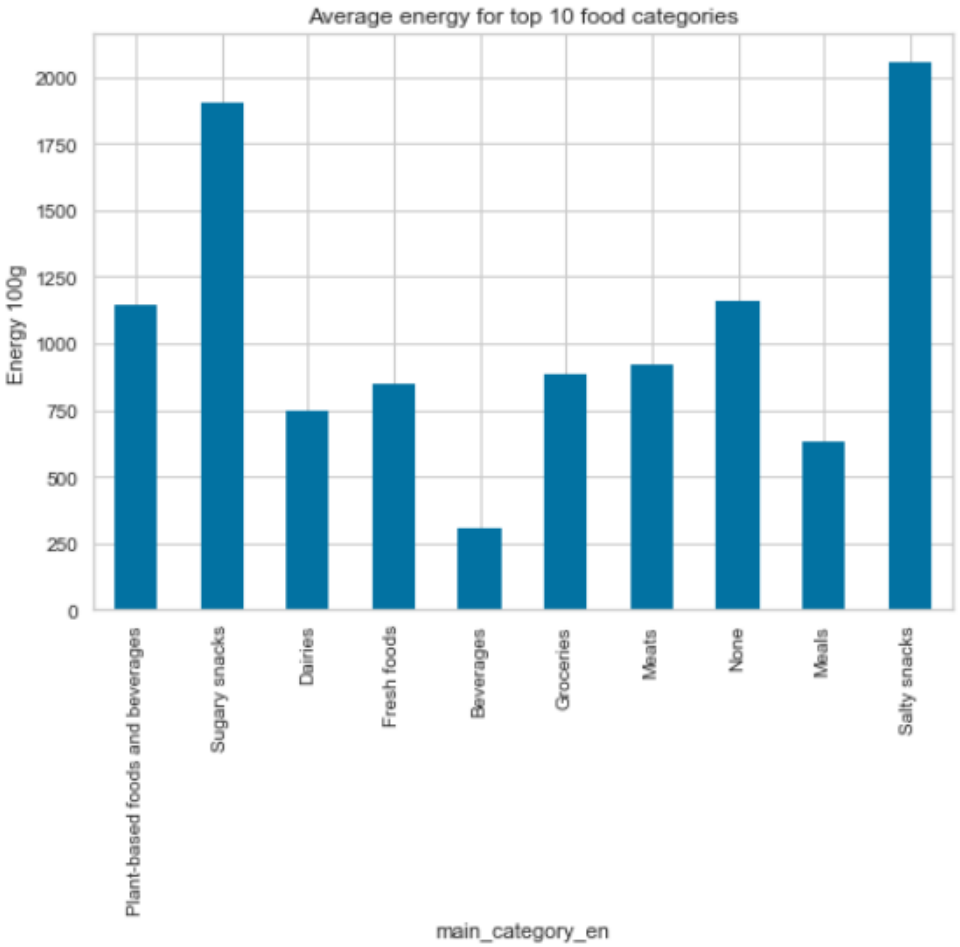


Figure 6: Food and their energy equivalent

6. Conclusion

In conclusion, kMeans is the best clustering algorithm among the ones that we used. It is fast, efficient and properly distributed the data points. We were able to successfully create a cluster based on the nutrients in food dataset and preform three different clustering algorithms on this cluster.

Disclaimer

Some of the images can be seen below due to the structuring of overleaf.

Also, this report is based on the 'Food.ipynb' Jupyter notebook. Not all the results, graphs and plot are presented in the report due to the page limit, however detailed analysis and answers to various research questions are reflected here.

Therefore, the report is meant as a summary and analysis on the most important conclusions. Please refer to the 'Food.ipynb' notebook to view other visualisations and analysis done.

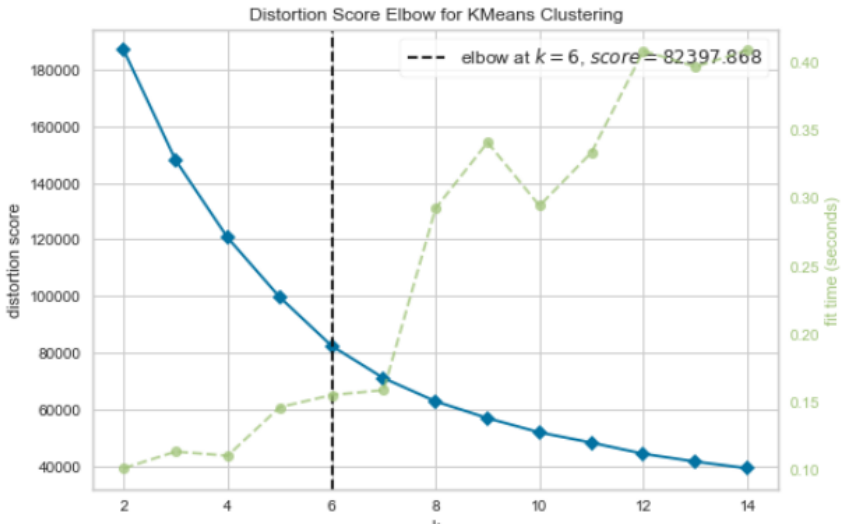


Figure 7: Optimal Value for K in KMeans

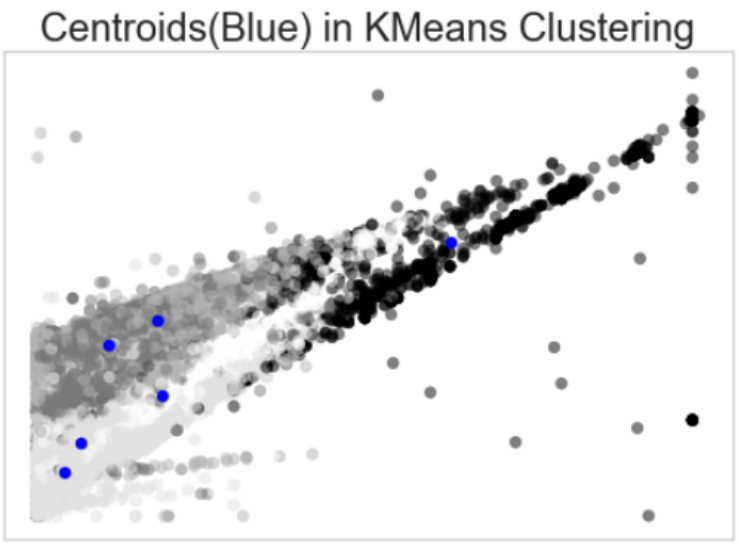
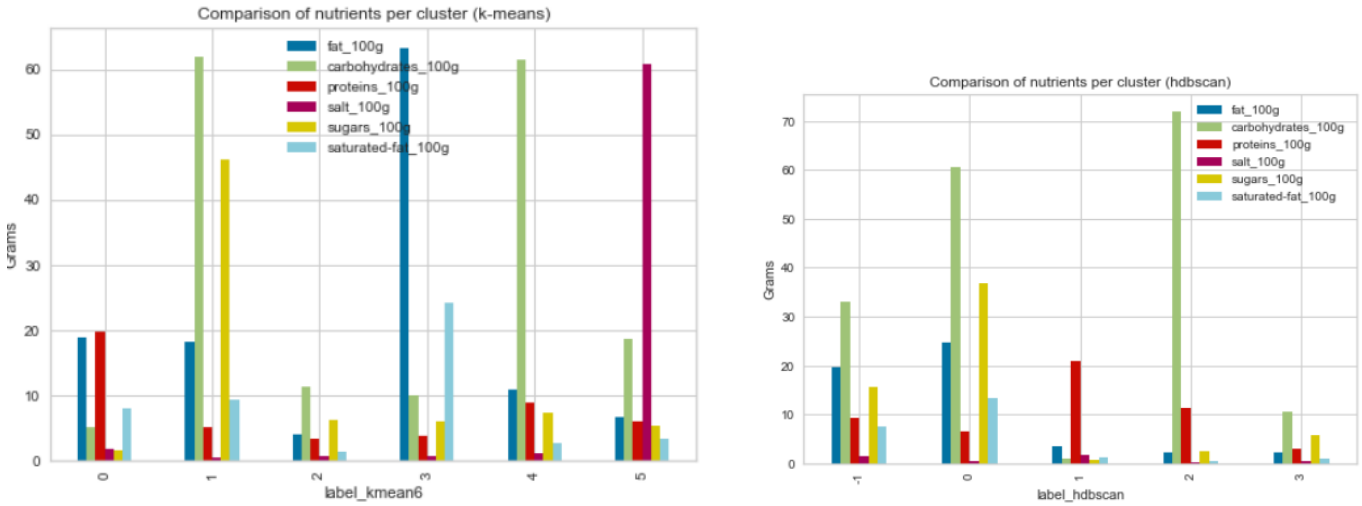
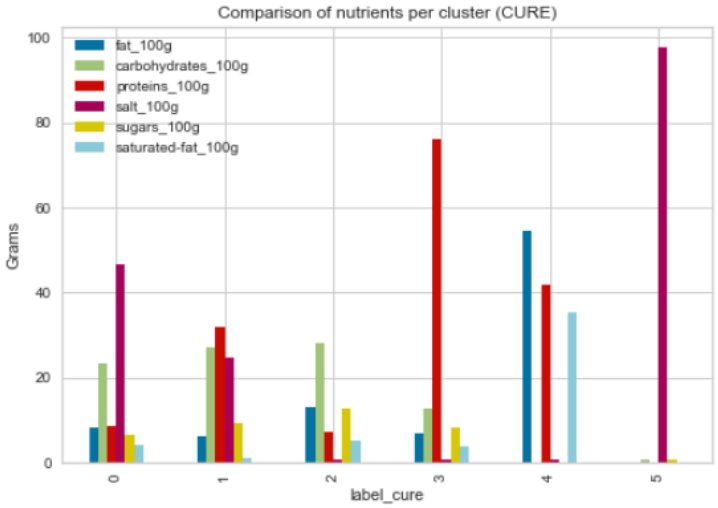


Figure 8: Centroids location in KMeans



(a) Comparison of nutrients per cluster in KMeans and HDBScan



(b) Comparison of nutrients per cluster in CURE

Figure 9: Outlier Display of variables

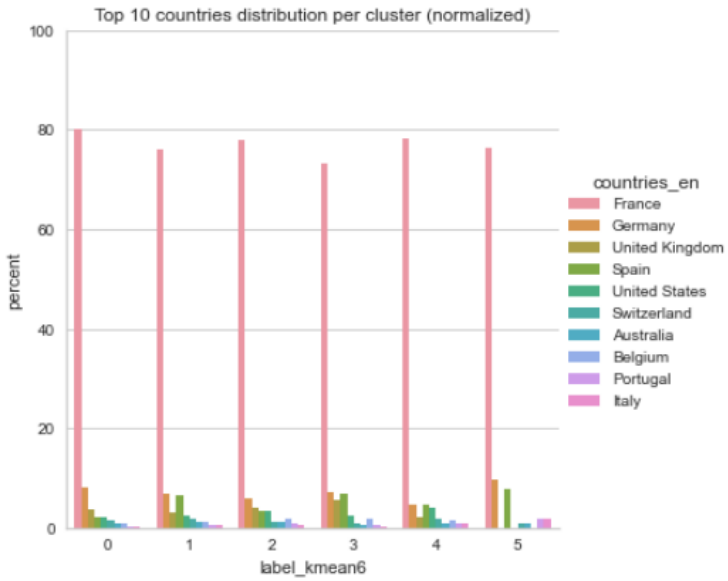


Figure 10: Normalisation of countries for kmeans clustering