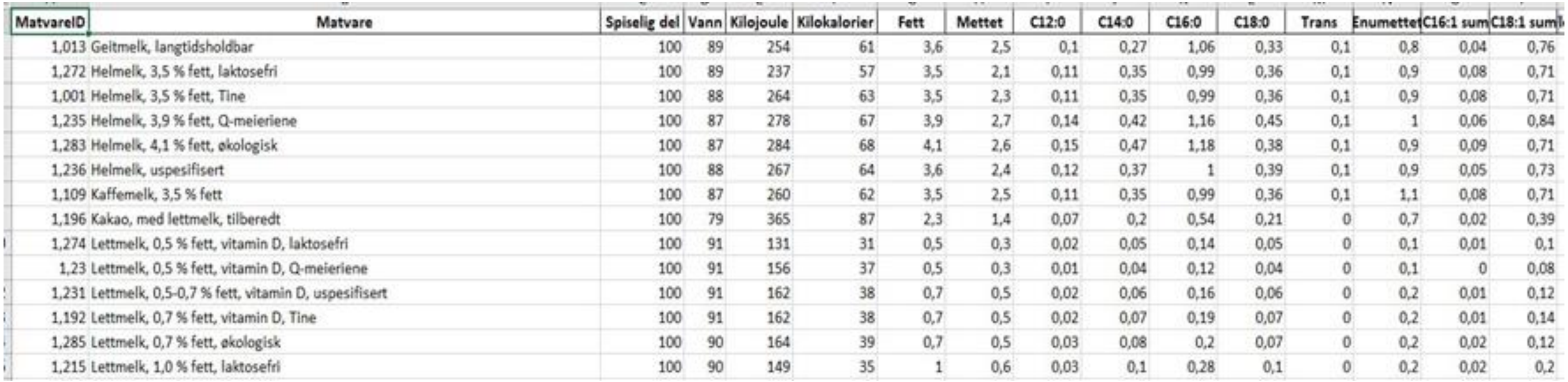
Find Similar Food Items

Problem

- Food is essential for life.
- Now a days, there are so many choices in terms of food items. And it becomes necessary to know the details about energy, nutrients & ingredients of the food items.
- In case, one particular food item isn’t available or not desired it’s beneficial to know the alternatives with the similar properties.
- Knowing the similar foods not only helps in finding alternatives, but also helps in selecting the food based on various factors like nutrition, carbon footprint, price etc.
- If people want to find healthier food items or foods with particular nutrition requirements , How do they find these alternatives ? Is there a database of all the food items (long list) they need to browse through ? Or can such a database be made available in a convenient manner to find the similar food items ?

Data

- For the purpose of finding similar food items, we use the Food table data provided by Norwegian Food Safety Authority. https://www.matportalen.no/verktoy/matvaretabellen/
- The food table shows values for the content of energy and nutrients in 100 grams of food for raw materials, products, prepared foods and dishes.
- Total 1878 food items are available in this table.
- For each food item(row), there are 57 fields (columns) containing energy and nutrients information.



MatvareID	Matvare	Spiselig del	Vann	Kilojoule	Kilokalorier	Fett	Mettet	C12:0	C14:0	C16:0	C18:0	Trans	Enumettet	C16:1 sum	C18:1 sum
1,013	Geitmelk, langtidsholdbar	100	89	254	61	3,6	2,5	0,1	0,27	1,06	0,33	0,1	0,8	0,04	0,76
1,272	Helmelk, 3,5 % fett, laktosefri	100	89	237	57	3,5	2,1	0,11	0,35	0,99	0,36	0,1	0,9	0,08	0,71
1,001	Helmelk, 3,5 % fett, Tine	100	88	264	63	3,5	2,3	0,11	0,35	0,99	0,36	0,1	0,9	0,08	0,71
1,235	Helmelk, 3,9 % fett, Q-meieriene	100	87	278	67	3,9	2,7	0,14	0,42	1,16	0,45	0,1	1	0,06	0,84
1,283	Helmelk, 4,1 % fett, økologisk	100	87	284	68	4,1	2,6	0,15	0,47	1,18	0,38	0,1	0,9	0,09	0,71
1,236	Helmelk, uspesifisert	100	88	267	64	3,6	2,4	0,12	0,37	1	0,39	0,1	0,9	0,05	0,73
1,109	Kaffemelk, 3,5 % fett	100	87	260	62	3,5	2,5	0,11	0,35	0,99	0,36	0,1	1,1	0,08	0,71
1,196	Kakao, med lettmelk, tilberedt	100	79	365	87	2,3	1,4	0,07	0,2	0,54	0,21	0	0,7	0,02	0,39
1,274	Lettmelk, 0,5 % fett, vitamin D, laktosefri	100	91	131	31	0,5	0,3	0,02	0,05	0,14	0,05	0	0,1	0,01	0,1
1,23	Lettmelk, 0,5 % fett, vitamin D, Q-meieriene	100	91	156	37	0,5	0,3	0,01	0,04	0,12	0,04	0	0,1	0	0,08
1,231	Lettmelk, 0,5-0,7 % fett, vitamin D, uspesifisert	100	91	162	38	0,7	0,5	0,02	0,06	0,16	0,06	0	0,2	0,01	0,12
1,192	Lettmelk, 0,7 % fett, vitamin D, Tine	100	91	162	38	0,7	0,5	0,02	0,07	0,19	0,07	0	0,2	0,01	0,14
1,285	Lettmelk, 0,7 % fett, økologisk	100	90	164	39	0,7	0,5	0,03	0,08	0,2	0,07	0	0,2	0,02	0,12
1,215	Lettmelk, 1,0 % fett, laktosefri	100	90	149	35	1	0,6	0,03	0,1	0,28	0,1	0	0,2	0,02	0,2

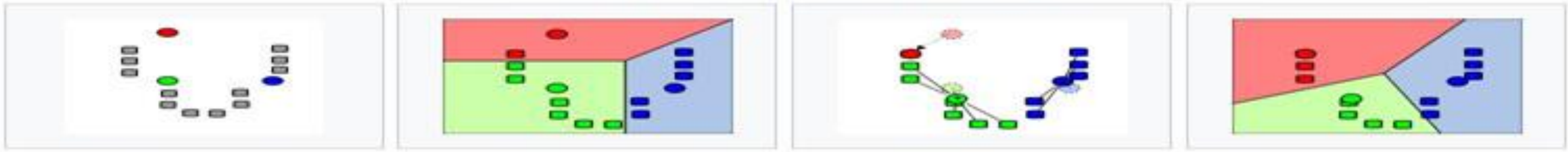
Data preprocessing

- Data contains rows for various food groups also, they're just group names with no information in other fields, hence those are removed.
- Row containing units of measurements is also not required for the analysis purpose, hence removed.
- We'll use the Row index to identify the 1878 food items. (Food ID and Food name columns are also removed, as they're not numeric data, we can merge them later if required)
- In columns, there are reference columns, which is again not relevant and non-numeric, hence that's also removed.
- Missing values can be filled with 0
- Data types of all the remaining fields has to be numeric.
- All the fields are scaled to [0,1] using Min-Max scaling.

Grouping Food Items using K-means clustering

- K-means is an unsupervised clustering algorithm, which allocates data points into groups based on similarity.

Demonstration of the standard algorithm



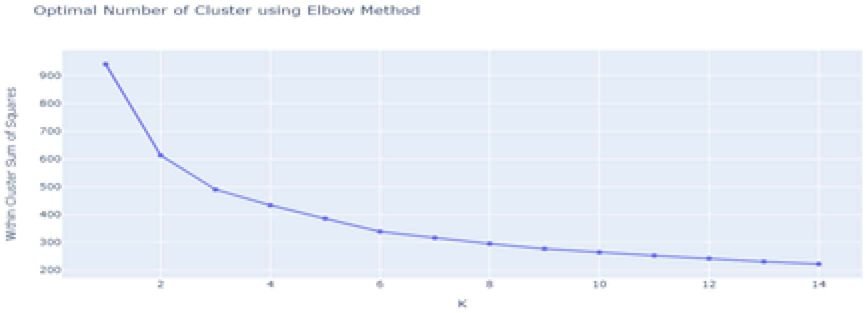
1. k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).

2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the k clusters becomes the new mean.

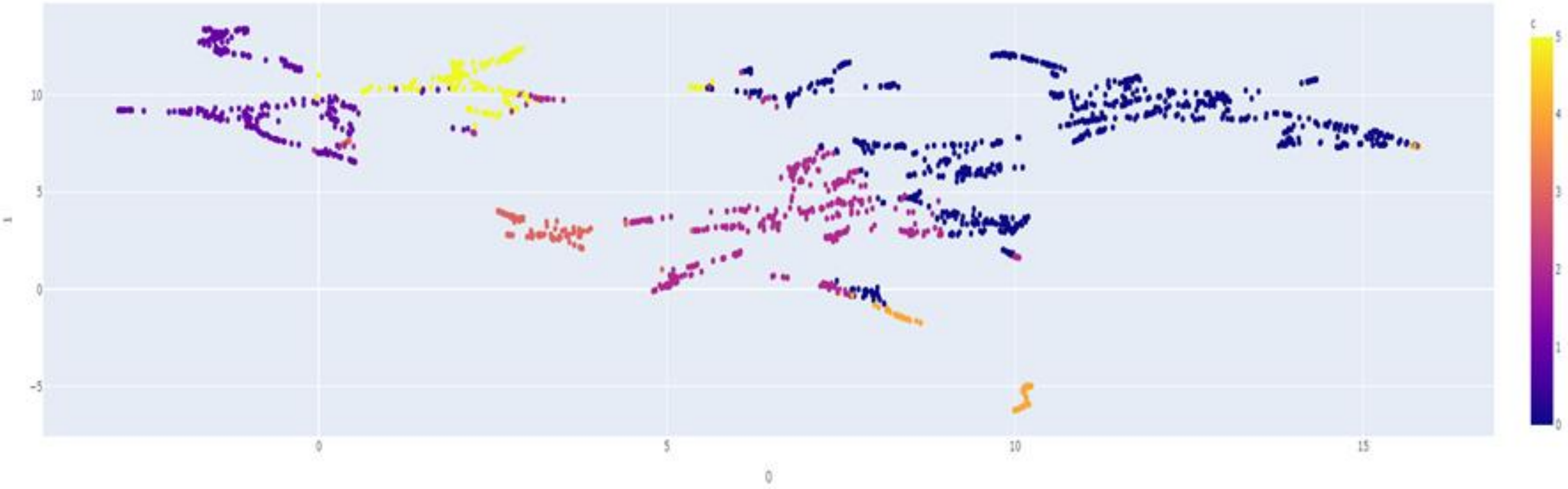
4. Steps 2 and 3 are repeated until convergence has been reached.

- K-means clustering is performed on preprocessed data of food items.
- **K = Number of clusters** is obtained using **Elbow method.** (K = 6 is the optimal no. of clusters as seen in the figure below)



Results

All Food Items when grouped/clustered into 6 clusters, can be visualized in 2D. (Plot below is obtained after dimensionality reduction 57 → 2, for visualization). Clusters look well separated. Items corresponding to the same cluster share similar properties hence can be used as alternatives or as options for the items in the same cluster.



Further Analysis

Doing further analysis using more information about:

1. Carbon footprint value
2. Eco friendly production