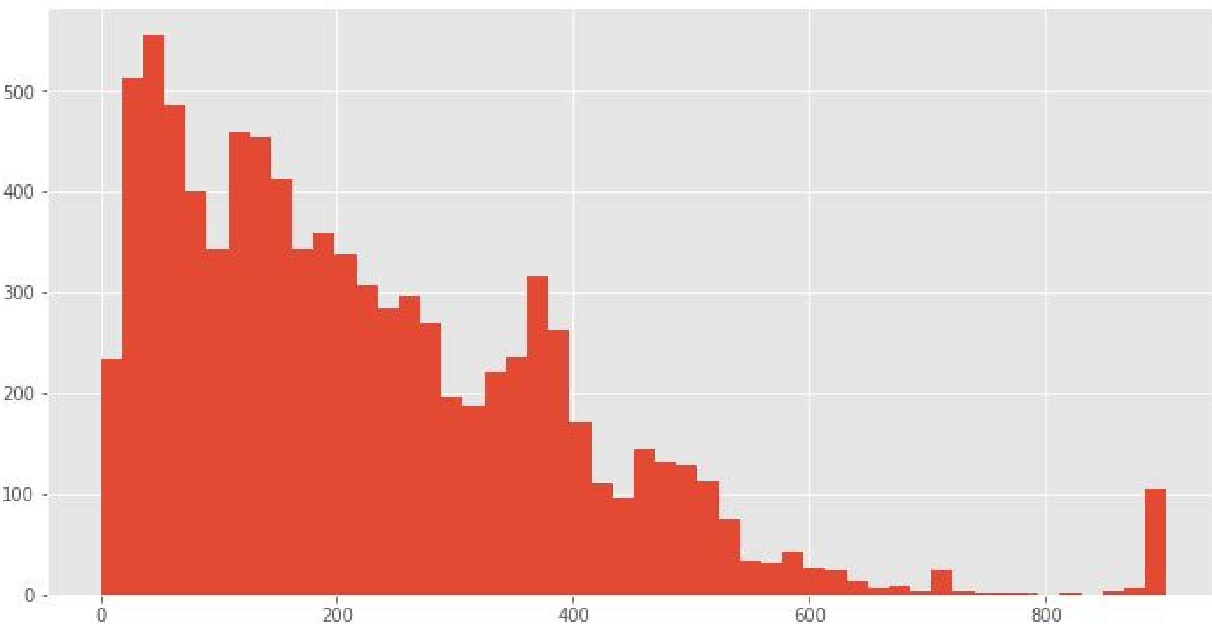
VU Visual Data Science WS 2018

Dzenan Hamzic BSc, TU Wien, 00327029

Lab Part 1

Task 1: Clustering

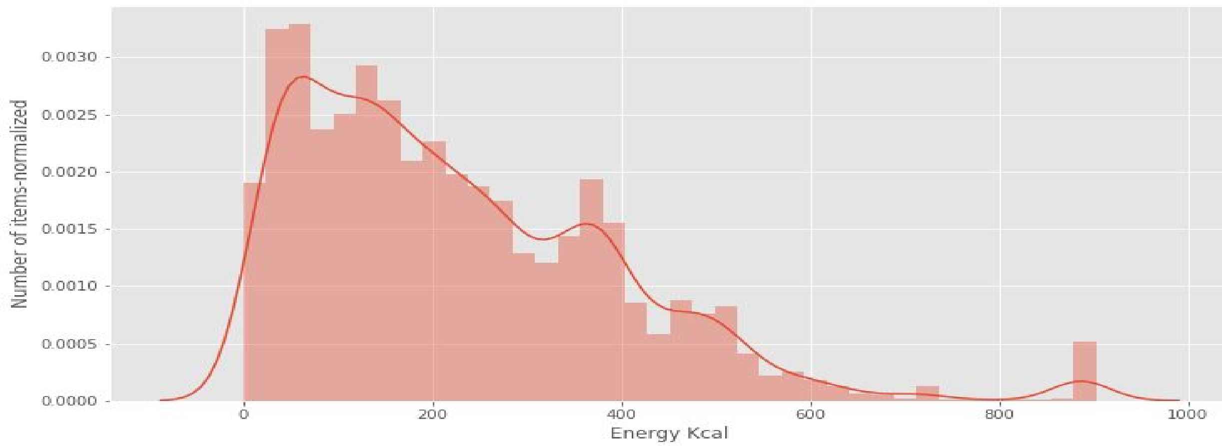
Visual Analysis



The number of bins(granularity) for histogram distribution plotting is crucial here. To small number of bins does not show real/appropriate underlying item distribution.

By looking at the histogram above, there are few clear groupings/clusters to be seen. By counting the peaks and the bins around them, I would say, there are 5 clusters in the data.

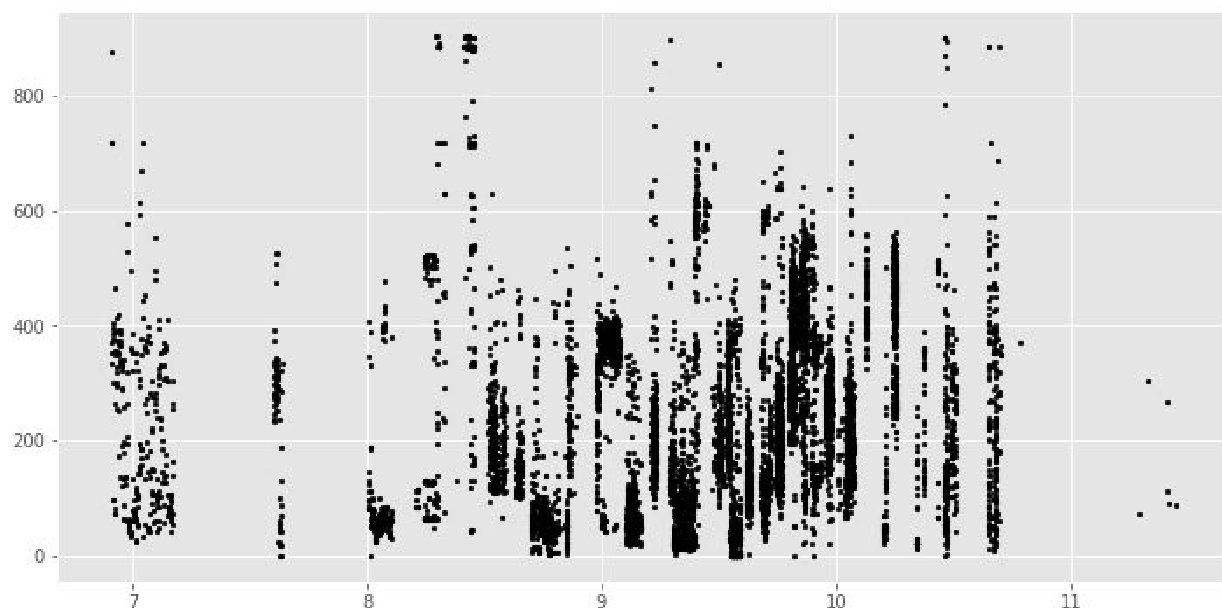
For the purpose of this subtask, I have also used seaborn library to visualize data clusters. The benefit of the seaborn's bar method is that it does not need apriori knowledge of bin number.



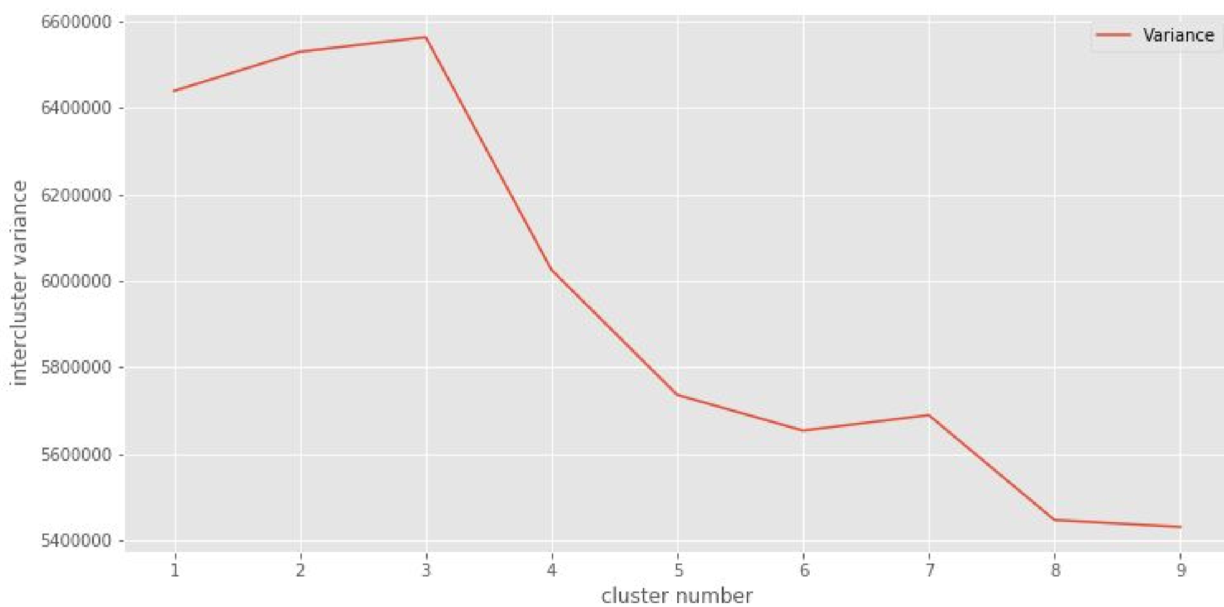
The smoothing line shows possible clusters nicely.

Statistical Analysis

In this section, I first tried to plot “Energ_Kcal”(y axis) vs. “NDB_No”(x axis) what made no sense with this data type.



For the purpose of statistical analysis I chose K-means clustering algorithm. This clustering algorithm is simple and very effective. The main drawback of the K-means is the prior knowledge of the number of clusters in the data. For this purpose, I decided to run K-means multiple times with stepwise increasing number of clusters, so called elbow method, in order to find optimal number of clusters in the data. The idea is to find minimal number of clusters that describe the most variance in the data.

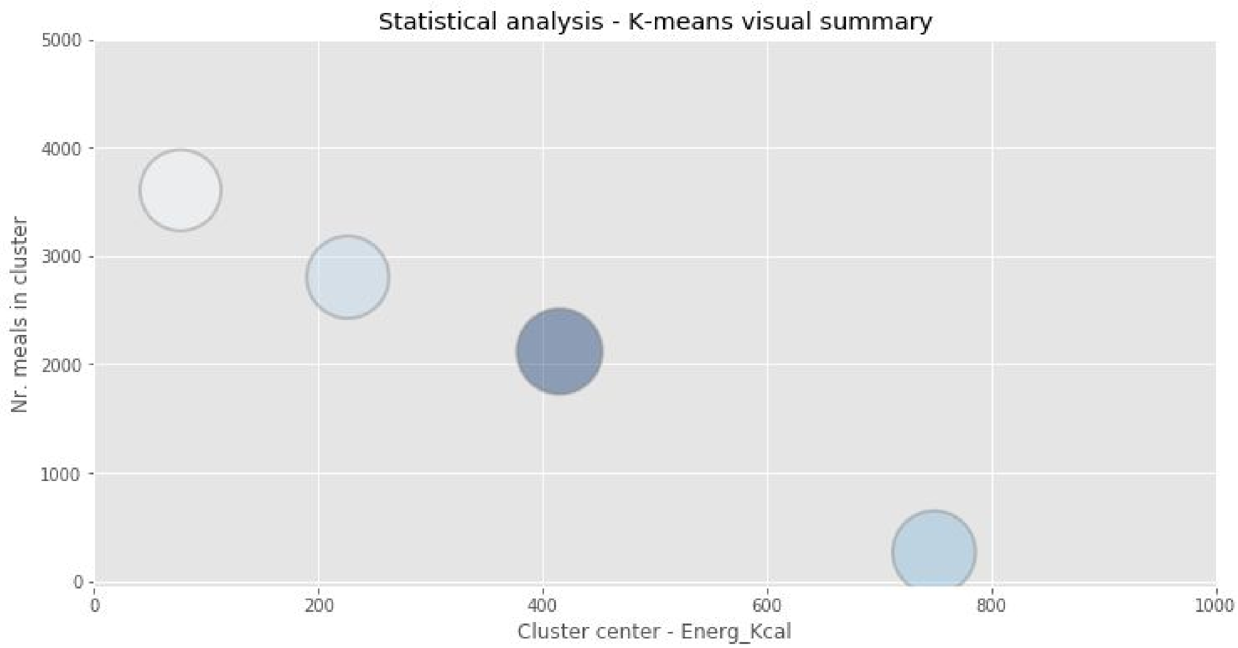


In this case, the elbow is obvious. The most gain in variance is the transition from 3 to 4 clusters. Therefore, for the purpose of this subtask/exercise, I will chose to do K-means clustering with 4 clusters.

Cluster summary.

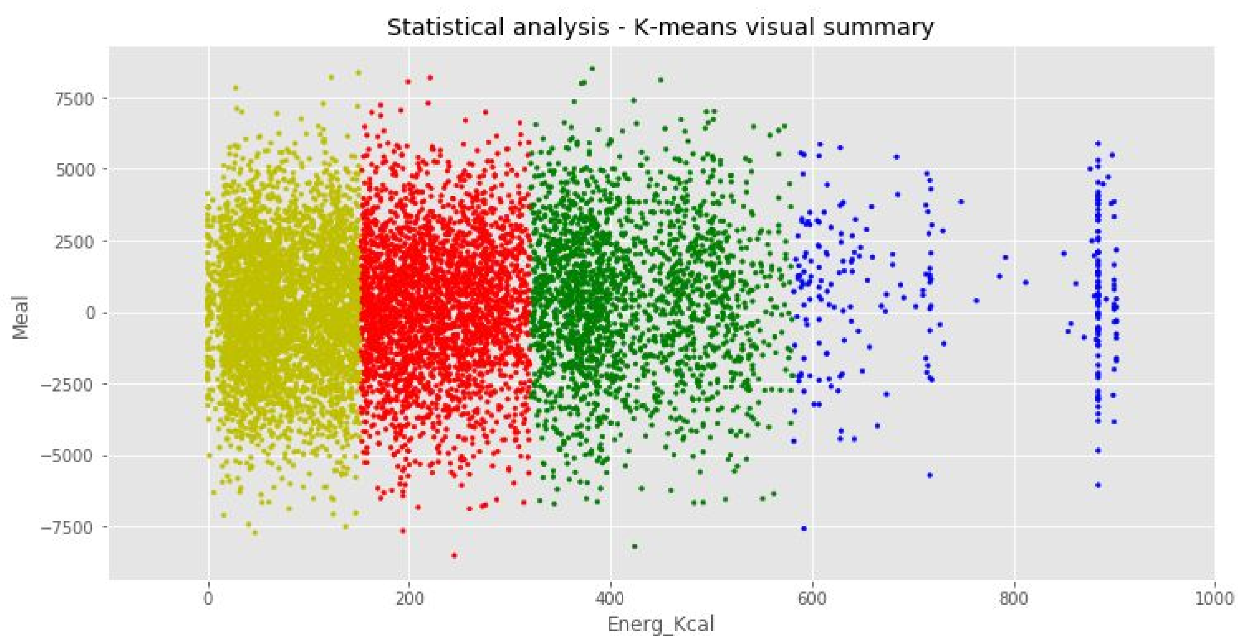
	0	1	2	3
center	226.28	415.00	748.84	77.22
nr.items	2803.00	2118.00	264.00	3605.00
variance	2408.57	2607.17	2455.16	2331.35

The four clusters have clearly different center values(means), roughly equal variances, and significantly unequal number if items that belong to that clusters. Based on this analysis, most of the meals 3605 belong to the low calorie cluster (number 0). Only 264 meals are in the "high calorie" cluster nr. 3. Below is the clustering visual summary.



This visual summary of clustering into 4 clusters points out the negative relationship between number meals and the number of high calorie meals in the data set. The bubble size indicates variance between items in a cluster.

For the sake of plotting all meals in one scatter plot, I have added new y dimension for every meal which value is random normally distributed around meals corresponding cluster center. This visual summary indicates that the points in the most right corner should be clustered separately.

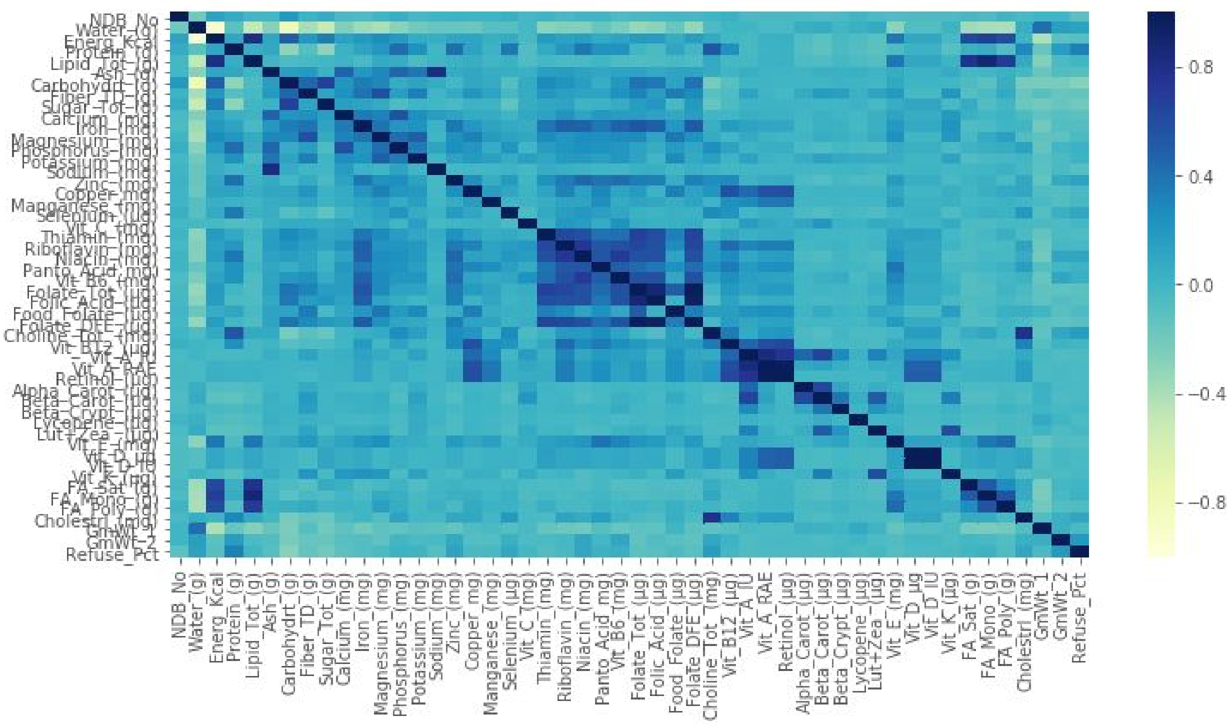


Statistical analysis indicates that 4 clusters are best option, but, visual analysis indicates that there should be one additional cluster with mean around 900. I think this point of visual analysis shows it's real power comparing to pure statistical approach. The choice of the cluster number from statistical analysis with K-means may be right, but the centers would have to be rearranged based on visual analysis. Visual part in this task is much easier to implement, but nonetheless, is crucial for data understanding.

Task 2: Correlations

Statistical analysis

I used pandas correlation function to visualize pairwise correlations in order to choose, interesting ones for me.



I decided to go with “Energ_Kcal” and “Water_(g)” which have strong negative correlation.

I inspected correlations with other variables using the firstly two chosen strongly negatively correlated variables.

"Water_(g)"		"Energy_Kcal"	
Energ_Kcal	-0.900554	Energ_Kcal	1.000000
Carbohydrt_(g)	-0.773920	Lipid_Tot_(g)	0.806677
Sugar_Tot_(g)	-0.506365	FA_Mono_(g)	0.691560
Lipid_Tot_(g)	-0.489781	FA_Sat_(g)	0.624444
FA_Poly_(g)	-0.405290	FA_Poly_(g)	0.607855
Magnesium_(mg)	-0.402719	Carbohydrt_(g)	0.493028
Fiber_TD_(g)	-0.394281	Vit_E_(mg)	0.370429
FA_Mono_(g)	-0.393146	Sugar_Tot_(g)	0.351313
FA_Sat_(g)	-0.366525	Magnesium_(mg)	0.266927
Iron_(mg)	-0.353255	Fiber_TD_(g)	0.204450
Folate_Tot_(µg)	-0.333919	Iron_(mg)	0.199372
Folate_DFE_(µg)	-0.324301	Phosphorus_(mg)	0.192235
Folic_Acid_(µg)	-0.294714	Folate_Tot_(µg)	0.186024

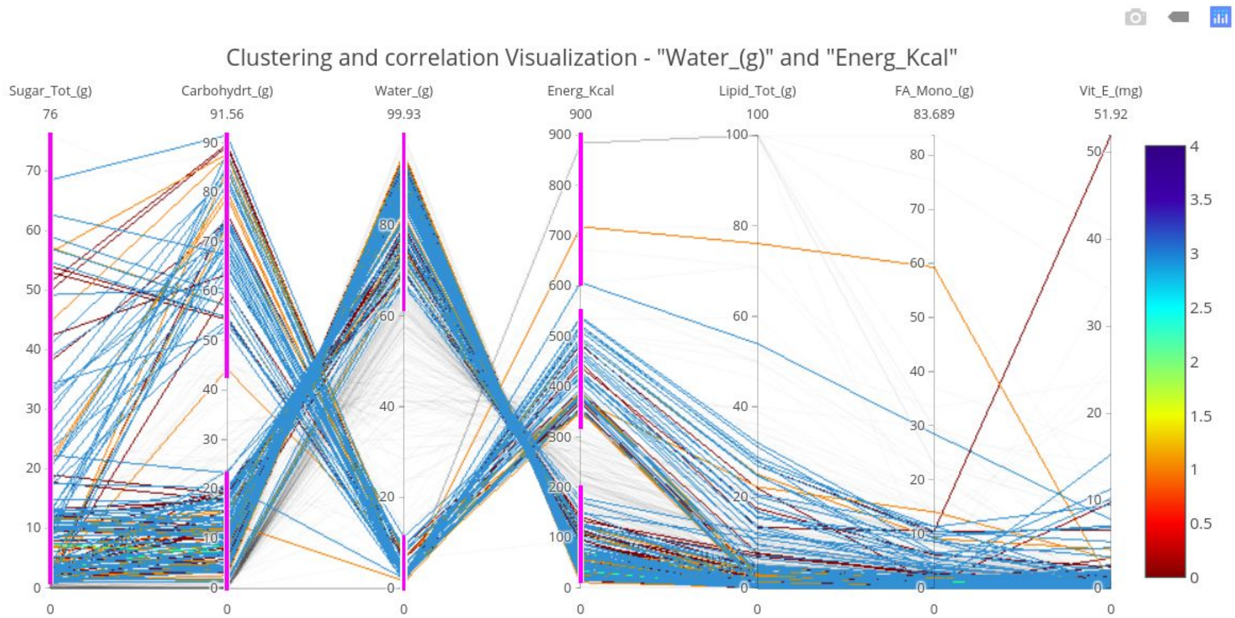
Based on this, I decided to take following additional variables.

##############################
Water_(g) -0.9 Energ_Kcal
Water_(g) -0.77 Carbohydrt_(g)
Water_(g) -0.51 Sugar_Tot_(g)

Energ_Kcal 0.81 Lipid_Tot_(g)
Energ_Kcal 0.37 Vit_E_(mg)
Energ_Kcal 0.69 FA_Mono_(g)
##############################

Visual Analysis

I have invested a lot of time in this task-subpart. The initial plot looked terrible. I tried many different approaches to get this plot “right”. I decided to color the items by the cluster number they are assigned to in K-means. Colour “blue” is the fourth cluster (third in the cluster summary table above) with lowest “Energ_Kcal” center value. I have used subsampling, to reduce plot overfitting, and brushing to clearly identify correlations in the plot below.



The parallel plot shows strong negative correlation between carbohydrates-water and water-energ_kcal. “Water” and “sugar” are slightly less negatively correlated. Energy-Kcal shows strong positive correlation with Lipids and less positive with FA_Mono and Vit_E.

The main difference in this subtask, between statistical and visual analytics, is that with statistical analysis you get only one number(correlation coefficient) which says nothing about the underlying data distribution and the amount of samples which it is based on. The visual analysis clarifies this issues and shows basically “what is happening” in the data and why. It is important to mention that statistical analysis in this part is much more easier to implement especially if the data set is huge. The variable order in the parallel plot is essential to be able to spot correlation types.

Task 3: Identify diffs between groups

In this subtask, I decided to go with "SUGAR FREE" meal type. I marked “SUGAR FREE” group as 1 in the dataset.

Statistical Analysis

The statistical analysis shows the following correlations between the two groups.

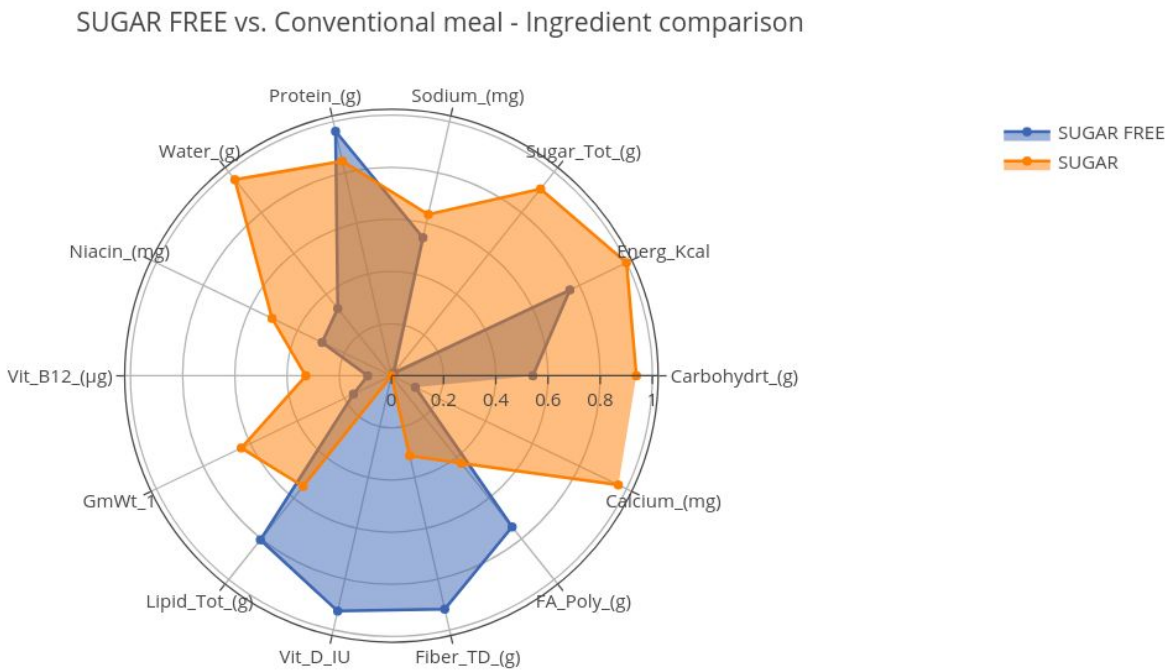
“SUGAR FREE” group correlations	
Carbohydrt_(g)	-0.534969
Energ_Kcal	-0.485929
Sodium_(mg)	-0.386612
Sugar_Tot_(g)	-0.349481
GmWt_1	-0.336831
Iron_(mg)	-0.314015
Protein_(g)	-0.310790
Thiamin_(mg)	-0.305649
Water_(g)	-0.292101
Niacin_(mg)	-0.284267
Lipid_Tot_(g)	-0.276312
Phosphorus_(mg)	-0.257447
Folate_Tot_(µg)	-0.246405
Zinc_(mg)	-0.241940
FA_Sat_(g)	-0.239497
FA_Mono_(g)	-0.236069
Ash_(g)	-0.229484
FA_Poly_(g)	-0.218444
Potassium_(mg)	-0.211624

Group mean values (chosen attributes)

	SUGAR_FREE	SUGAR
Carbohydrt_(g)	16.807143	29.102468
Energ_Kcal	117.675325	155.012987
Sugar_Tot_(g)	0.257013	15.789351
Sodium_(mg)	84.233766	98.441558
Protein_(g)	1.491039	1.309610
Water_(g)	10.227662	29.835325
Niacin_(mg)	0.914610	1.573870
Vit_B12_(µg)	0.094026	0.339221
GmWt_1	25.048052	98.957792
Lipid_Tot_(g)	6.239481	4.207013
Vit_D_IU	0.220779	0.000000
Fiber_TD_(g)	1.779221	0.609091
FA_Poly_(g)	1.916390	1.110013
Calcium_(mg)	1.051948	9.974026

Visual analysis

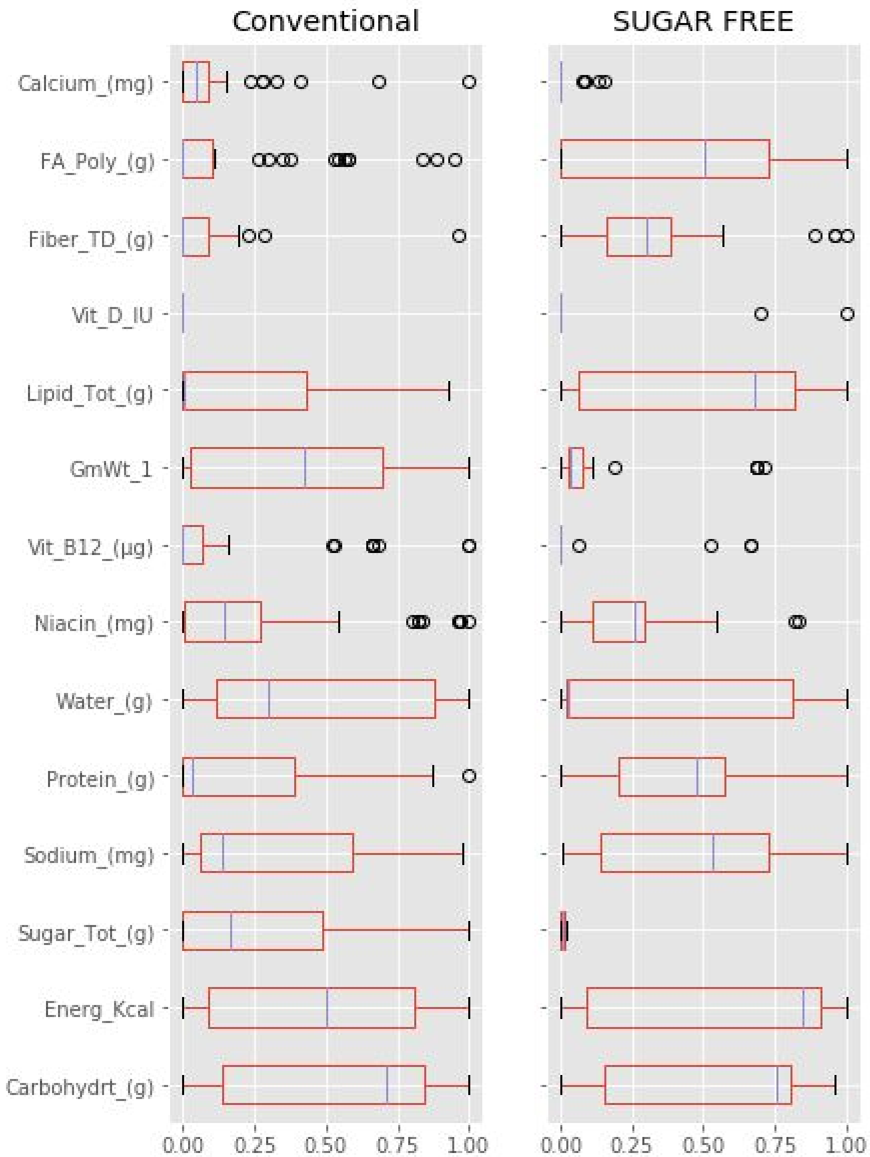
This subtask was a bit tricky since the attribute mean values are not on the same scale. Furthermore, when the variable scales are aligned, the ratios are too different for the plot to make sense. Therefore, I have done some manual, after variable values normalization, scaling on variables which values were too small comparing to others I choose to plot. The radar plot below shows group relative differences on respective variables. Please note that scales between attributes are different and are not relative in this case. The idea was just to show between group difference on chosen attributes.



The human visual system does not perceive the group differences from the statistical analysis as from radar plot. Interpreting the differences from the radar plot is much easier and more convenient. The statistical analysis shows pure numbers from which is hard for a human to identify a set of multiple attributes which are clearly more represented in one or another group.

To conclude, I would say that both, statistical and visual, parts are important and crucial in good data experiment. They supplement each other, fill in each other weaknesses, and are nonseparable. I would dare to say that good visual data analysis is harder to implement, but on the other hand, gives much more data insight.

Since I enjoyed this task very much, I decided to also do the list of box plots.



The whole experiment/notebook (without plotly visualizations) can be found at https://bit.ly/2Lm9Nxx
With Plotly viz : https://bit.ly/2GlDFeI