Nutrition and non-communicable disease risk factors

A talk by Simon Müller

for the lecture Introduction to Data Science

- **Introduction**

- **Data & Methods**
- *Data Sources*
- *Principal Component Analysis*
- *Clustering*

- **Results**
- *Scatter plots*
- *K-Means clustering*
- *Correlations*

- **Conclusion**

Introduction and Goals

- Combination of world nutrition, health indicator and GNI per capita datasets
- Apply Principal Component Analysis and clustering to find interesting patterns in the data
- Quantify the correlation between different food groups, health indicators and income

Food and Agriculture Organization of the United Nations

Data Sources

Nutrition database:

- FAO.org
- Food Balance Sheets (1961 – 2013)
- Item of interest: Food supply (kcal/capita/day)

The data is aggregated by food groups.

Data Sources

Health indicator database:

- WHO.int
- Global Health Observatory data
- Non-communicable diseases
- Common risk factors include unhealthy diet, overweight, raised blood pressure, blood sugar and cholesterol

Data Sources

Given percentage of population that is affected by:

- Raised fasting blood glucose ($\geq$ 7 mmol/L or on medication)
- Raised total cholesterol ($\geq$ 5 mmol/L)
- Raised blood pressure (SBP $\geq$ 140 or DBP$\geq$ 90), age-standardized
- Overweight (body mass index $\geq$ 25), age-standardized

Also:

- HALE: Healthy Life Expectancy Index

Data Sources

GNI per capita to estimate income:

- data.worldbank.org
- Latest (2015) values in USD
- This dataset is only used for the correlation part of this talk

Principal Component Analysis

Reminders:

- PCA emphasizes variation and can be used to reduce the dimensionality of a dataset
- It involves finding linear combinations of a set of variables that has maximum variance and removing its effect

In my case:

- Up to 20 dimensional dataset
- PCA allows to visualise the data in two or three dimensions

Clustering

- Flat clustering with the K-Means algorithm
- Metric used: Distance between points
- Parameter: Number of clusters (n = 6)
- Six regions: Africa, Americas, Europe, Eastern Mediterranean, South-East Asia, Western Pacific

- **Introduction**

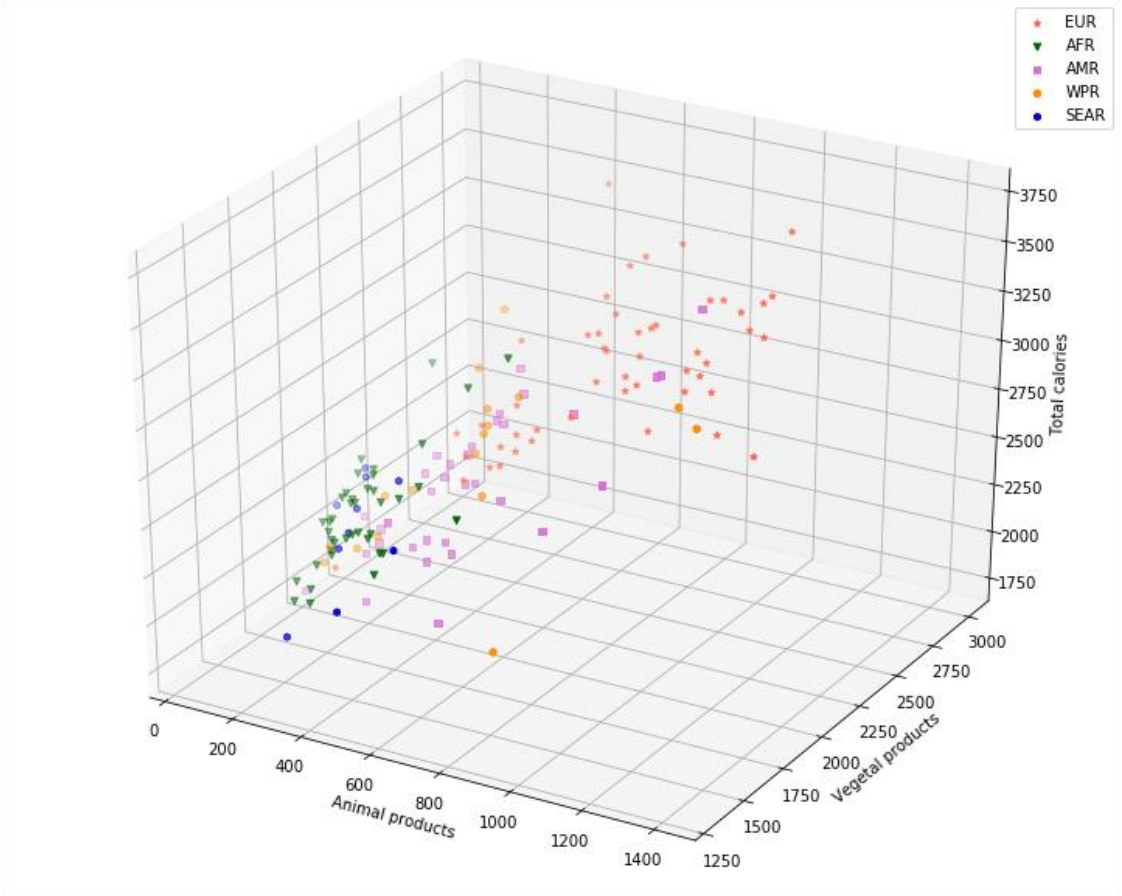
- **Data & Methods**
- *Data Sources*
- *Principal Component Analysis*
- *Clustering*

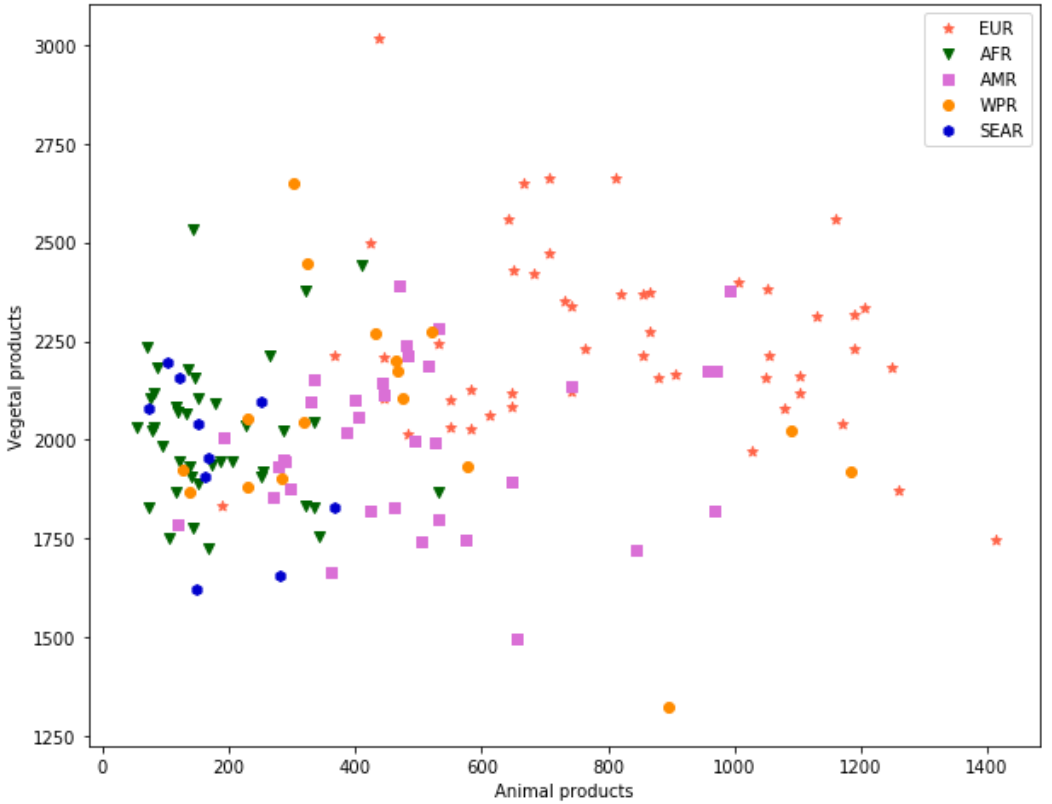
- **Results**
- *Scatter plots*
- *K-Means clustering*
- *Correlations*

- **Conclusion**

Scatter plots

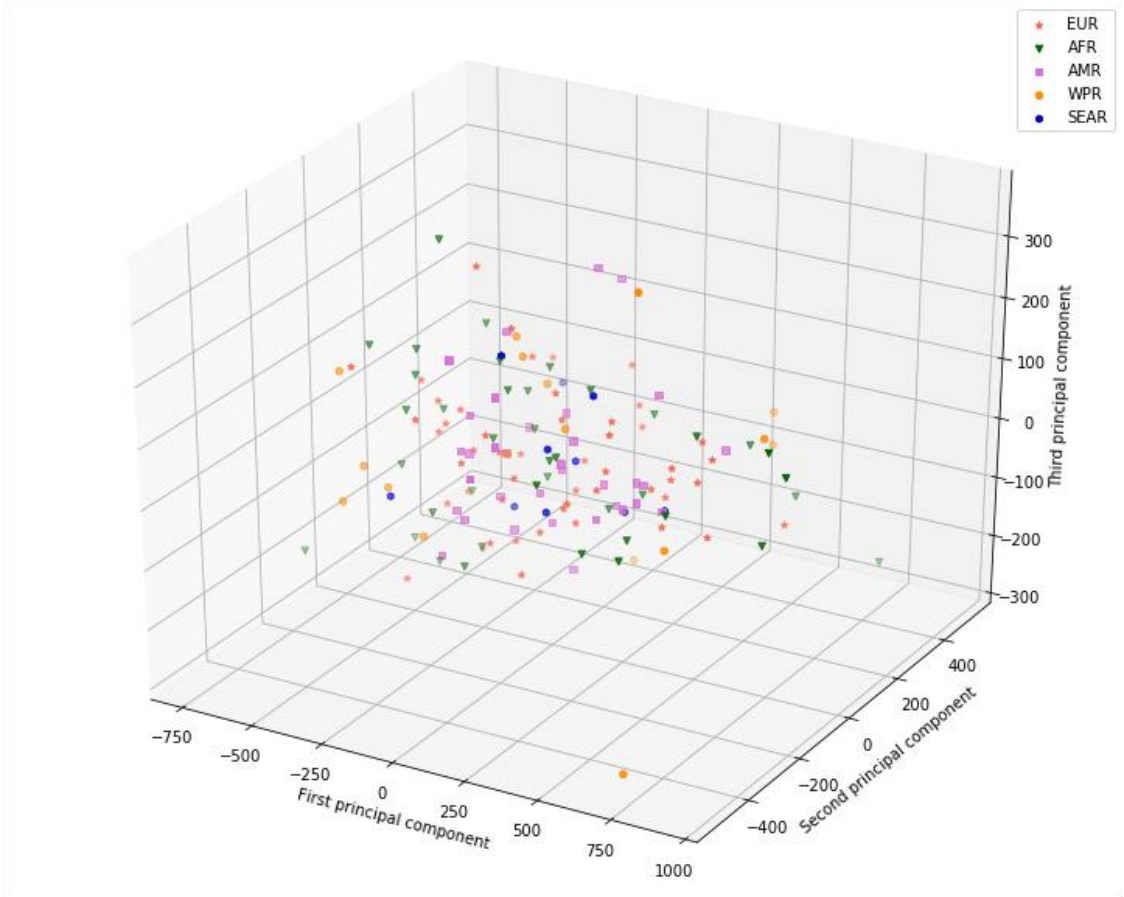
- Let's start with something we can plot without manipulation

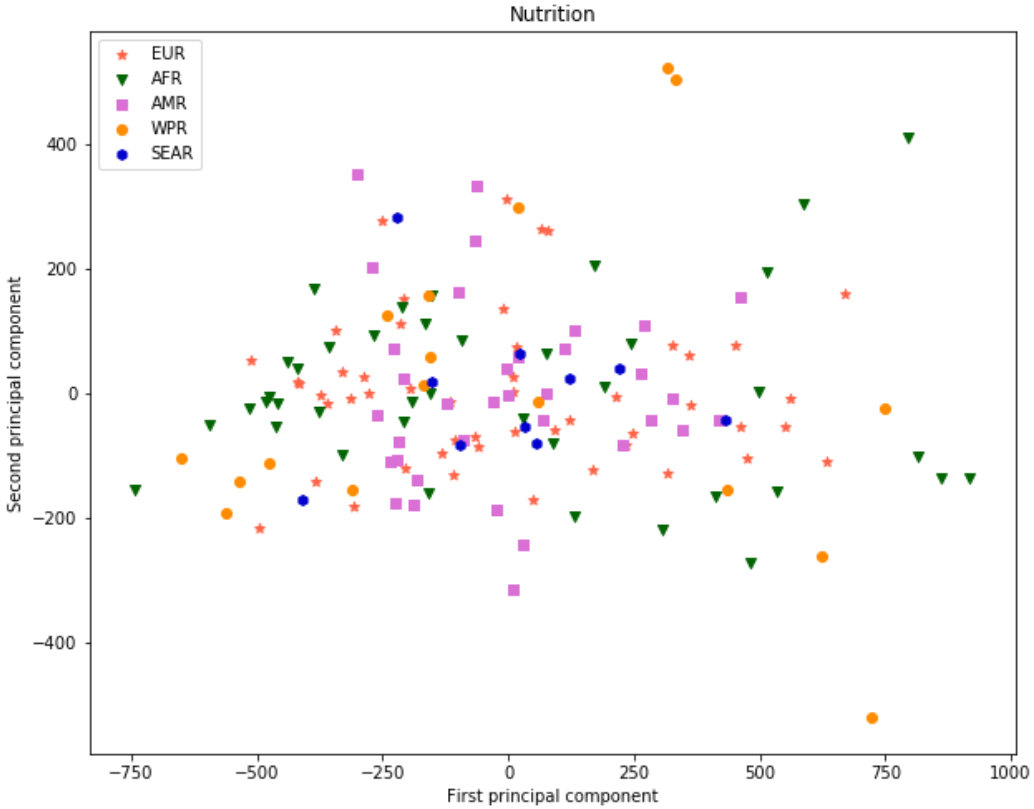




Scatter plots

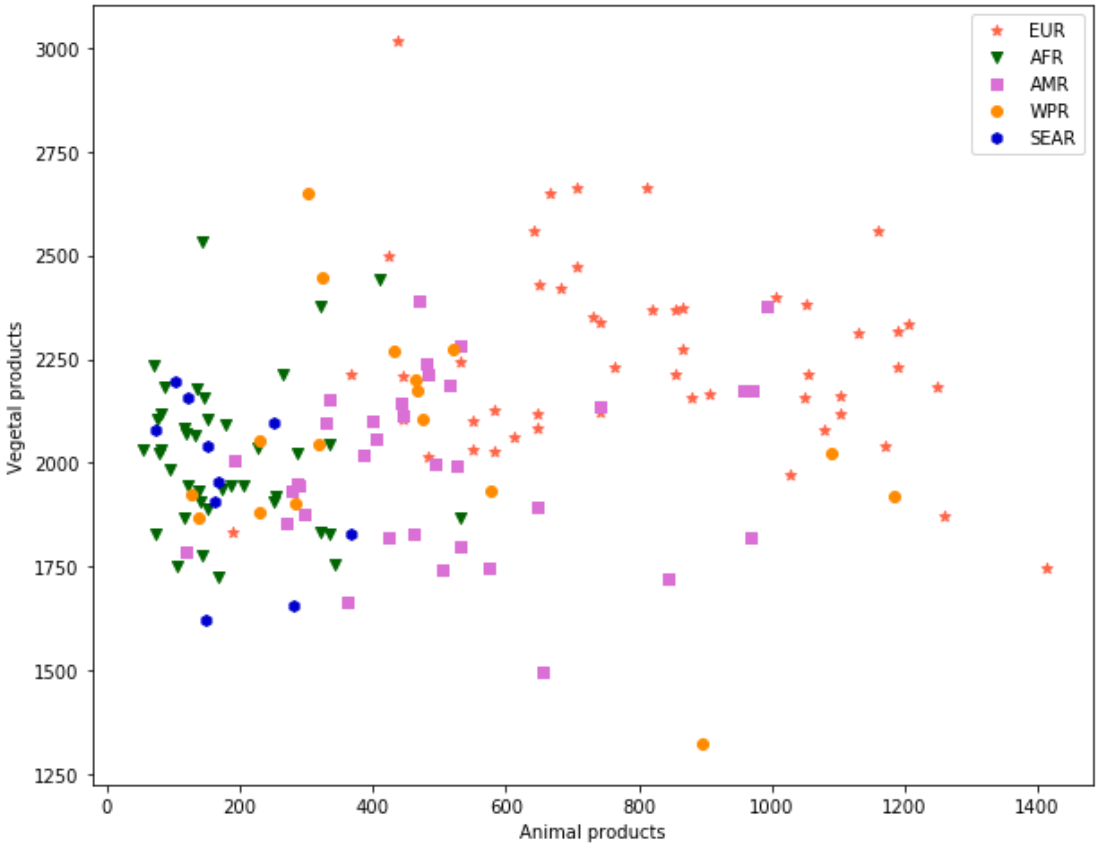
- PCA on the full nutrition data set

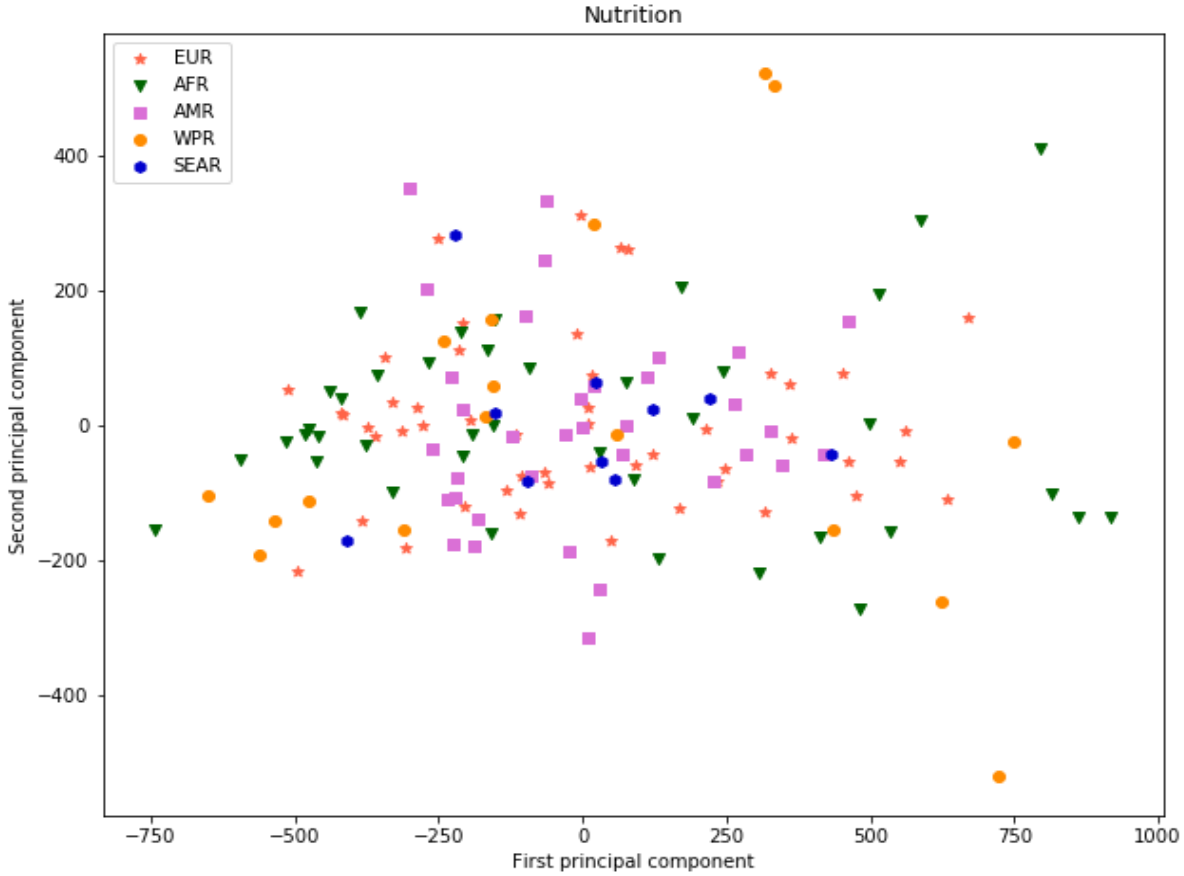




Scatter plots

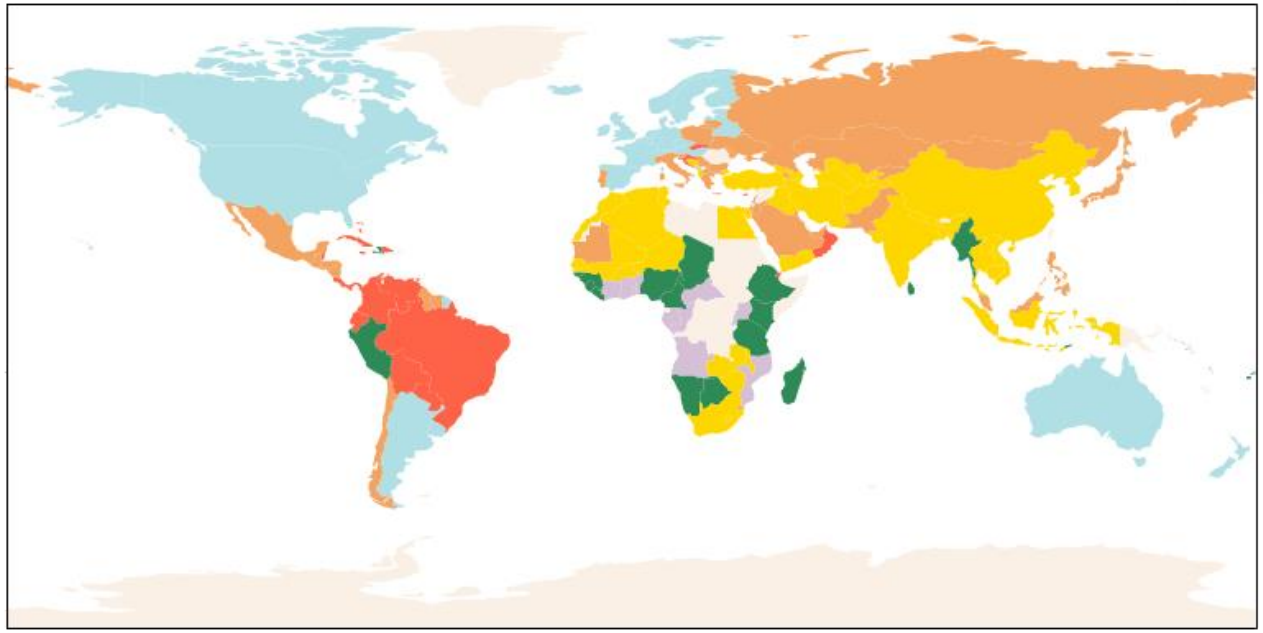
- Comparison

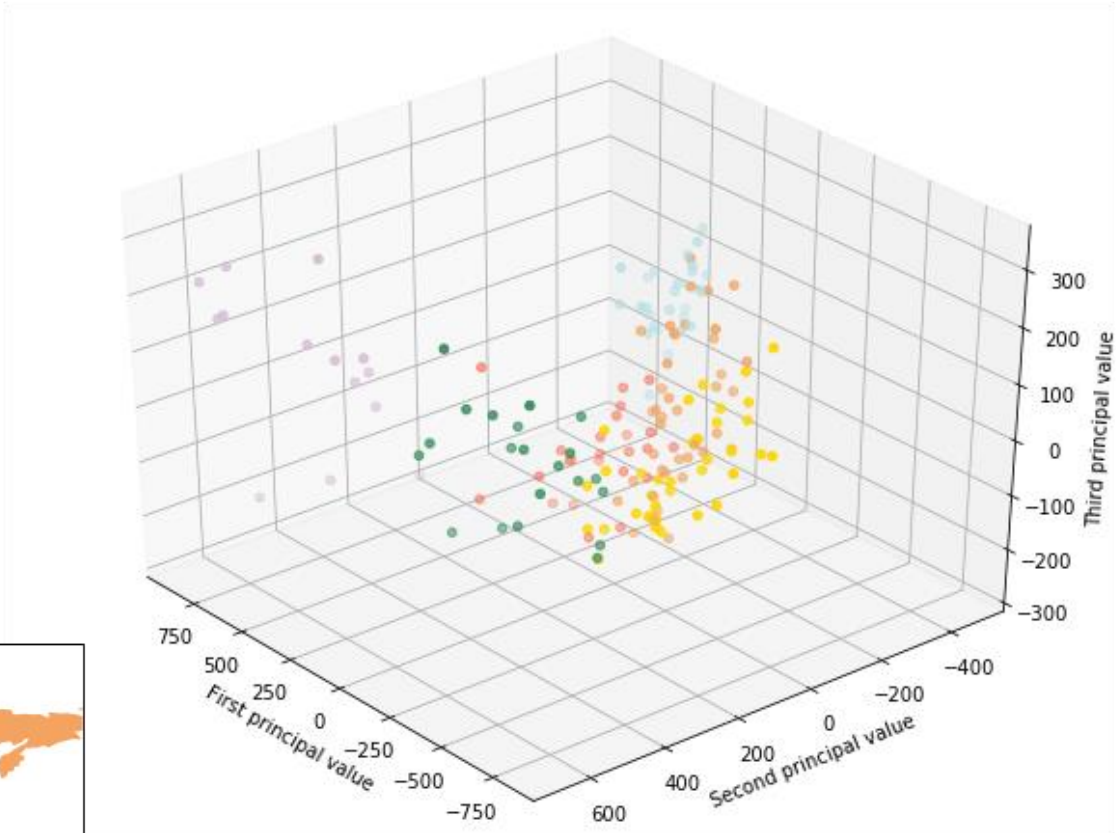




K-Means clustering

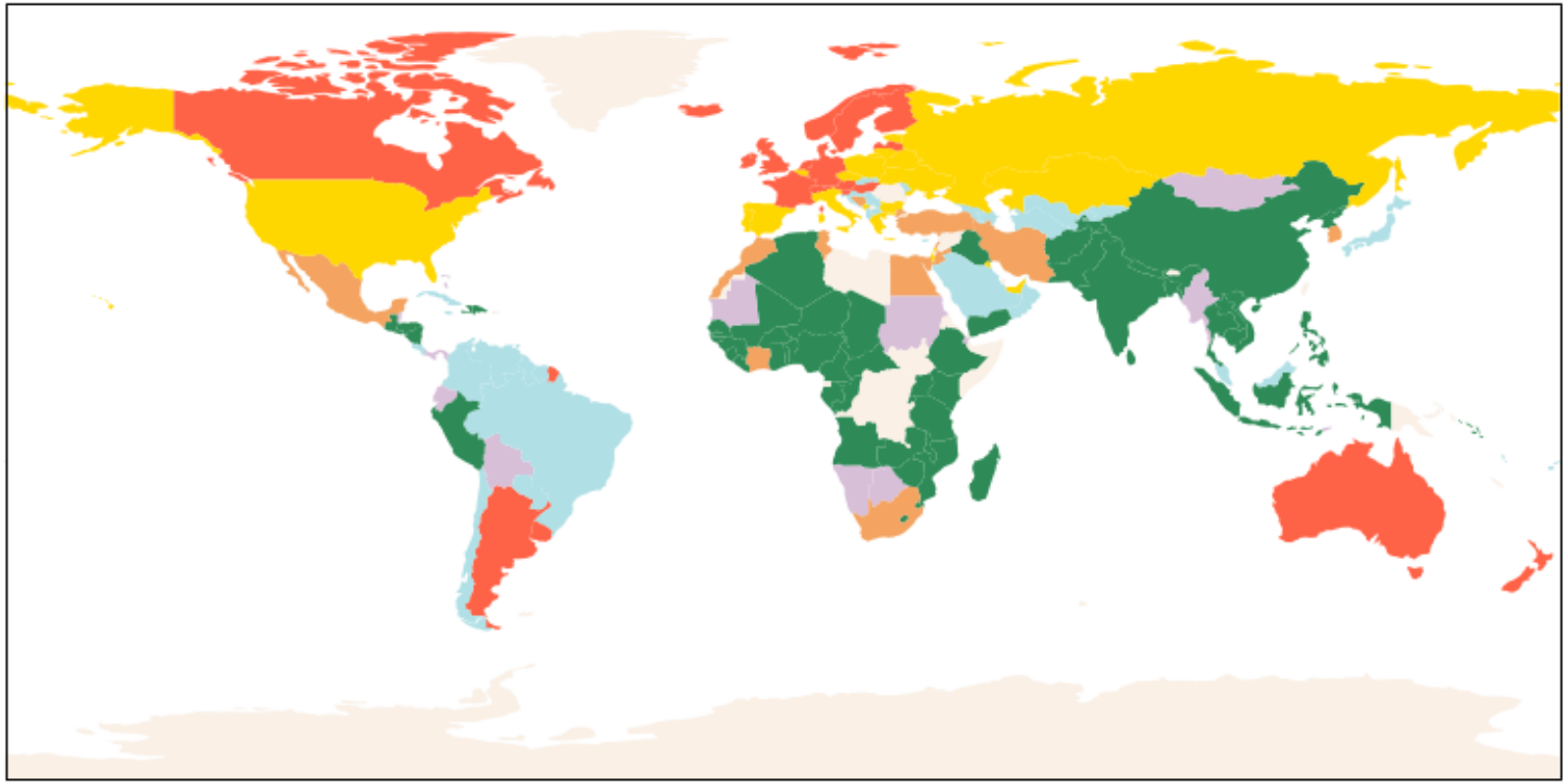
- Initiate K-Means clustering with n = 6 clusters (number of world regions)
- Use PCA on the full data to reduce it to three dimensions





K-Means clustering

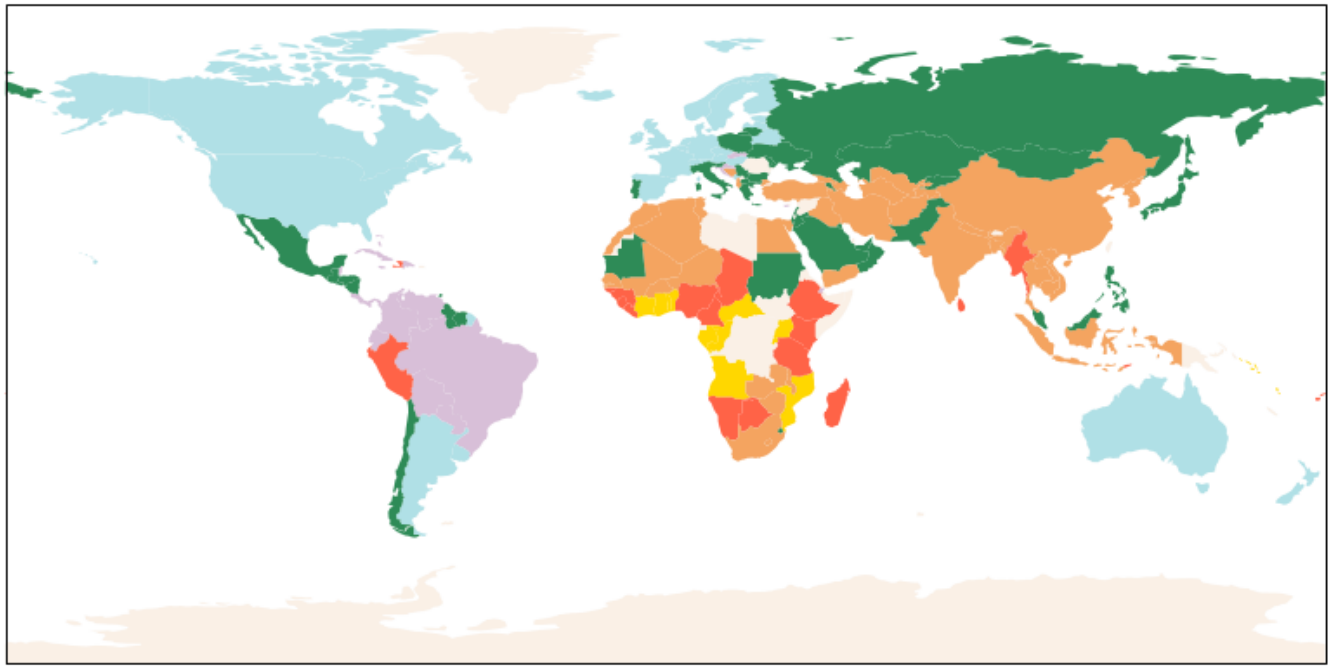
- Clustering on the coarsely aggregated nutrition data

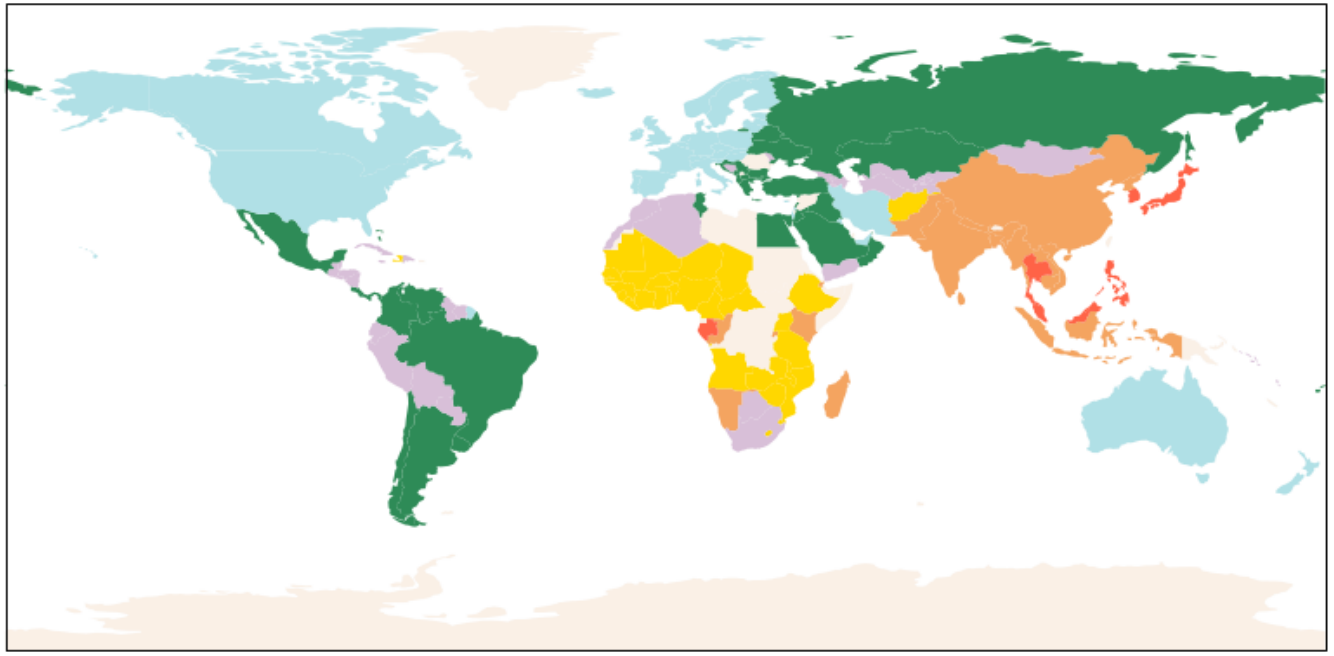


K-Means clustering

- We can use clustering on n-dimensional data and then use the world map to visualise the result

- Next slide: Clustering based on only the nutrition data versus the health indicator data





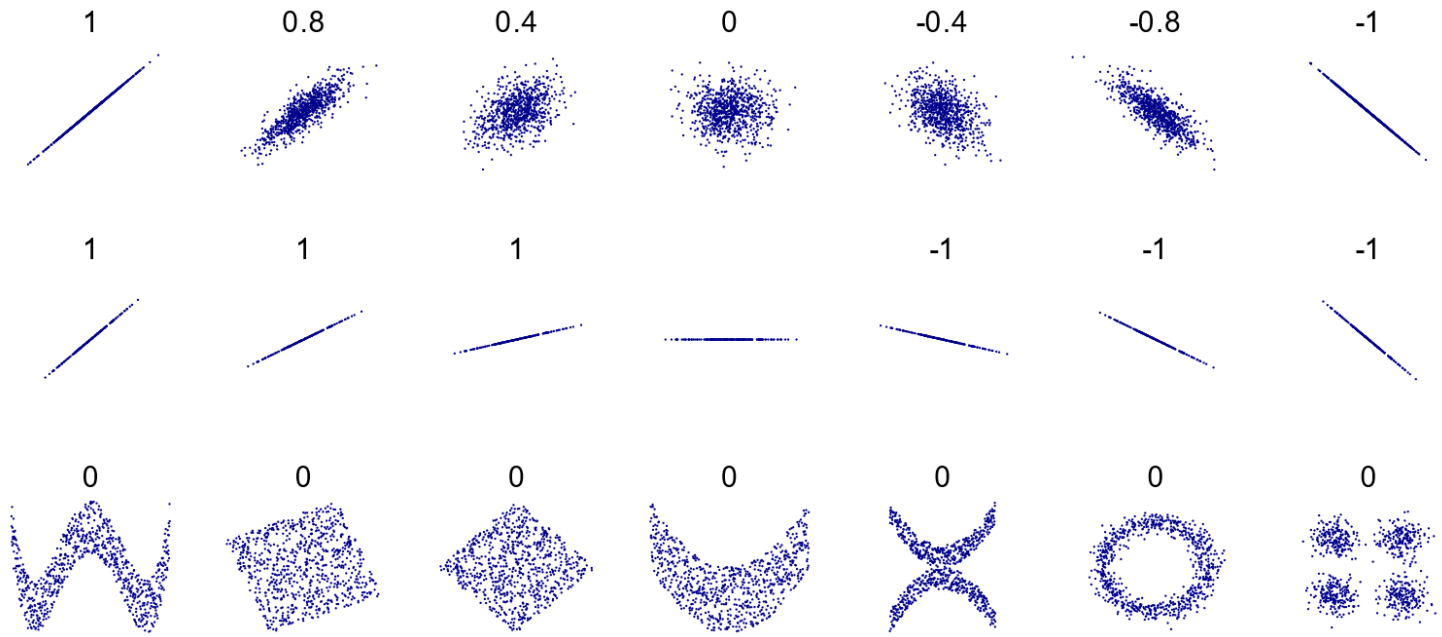
Correlations

We can use the Pearson Correlation Coefficient (PCC) to try and quantify correlations.

Generally used interpretation:

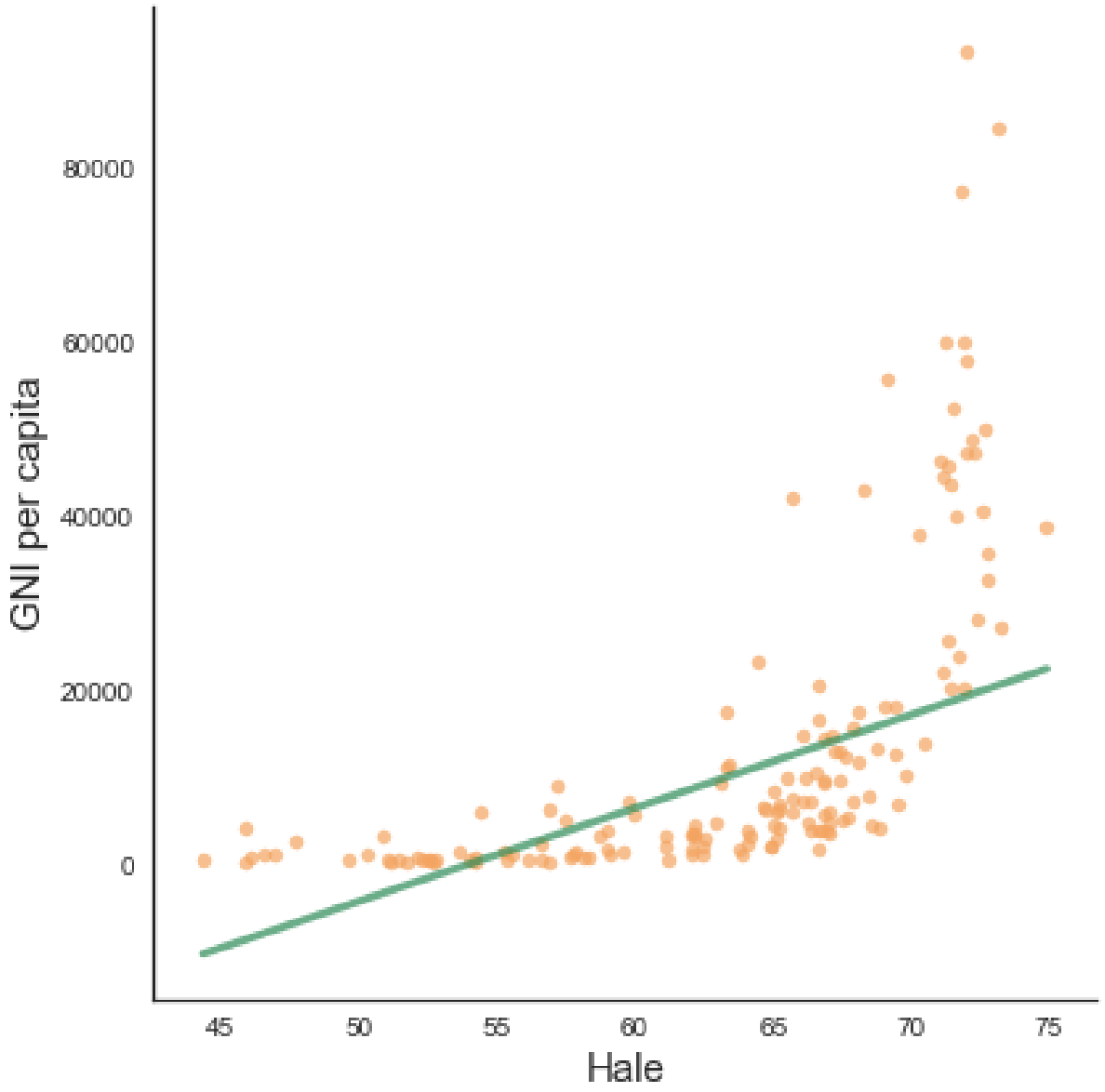
- 0.1 < r <0.3: Small
- 0.3 < r < 0.5 : Moderate
- r > 0.5: Strong

Caution: Only valid when considering linear relations

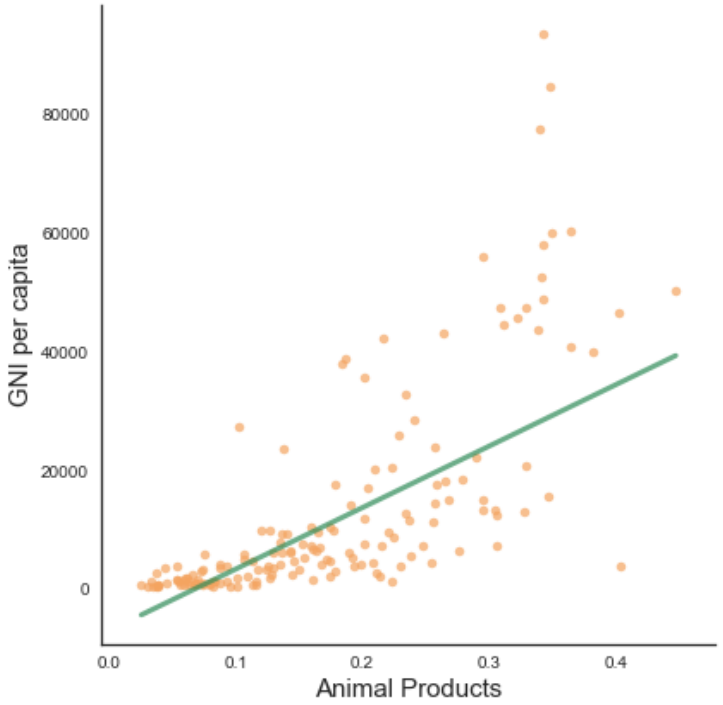


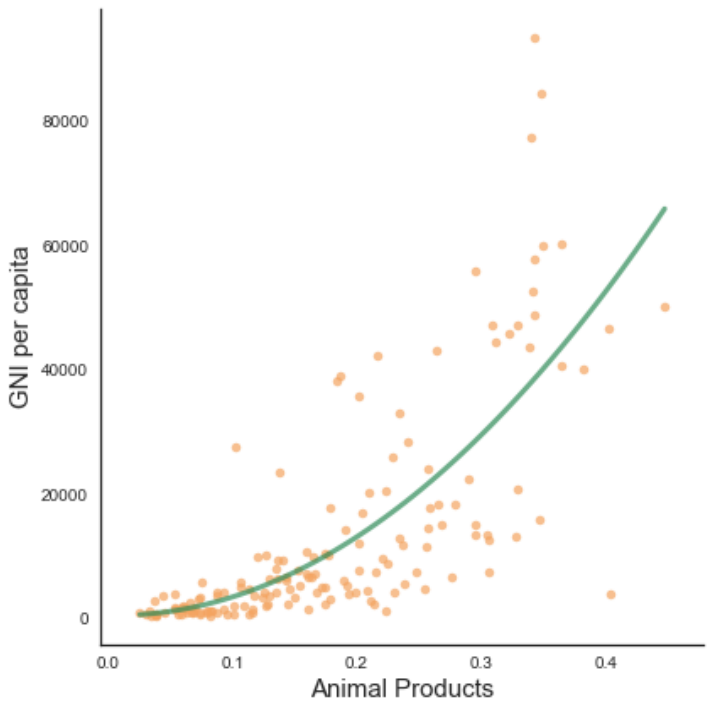
Credits: Wikipedia

	GNI per capita
Blood pressure	0.05
Blood glucose	-0.09
Cholesterol	0.75
BMI	0.48
HALE	0.63

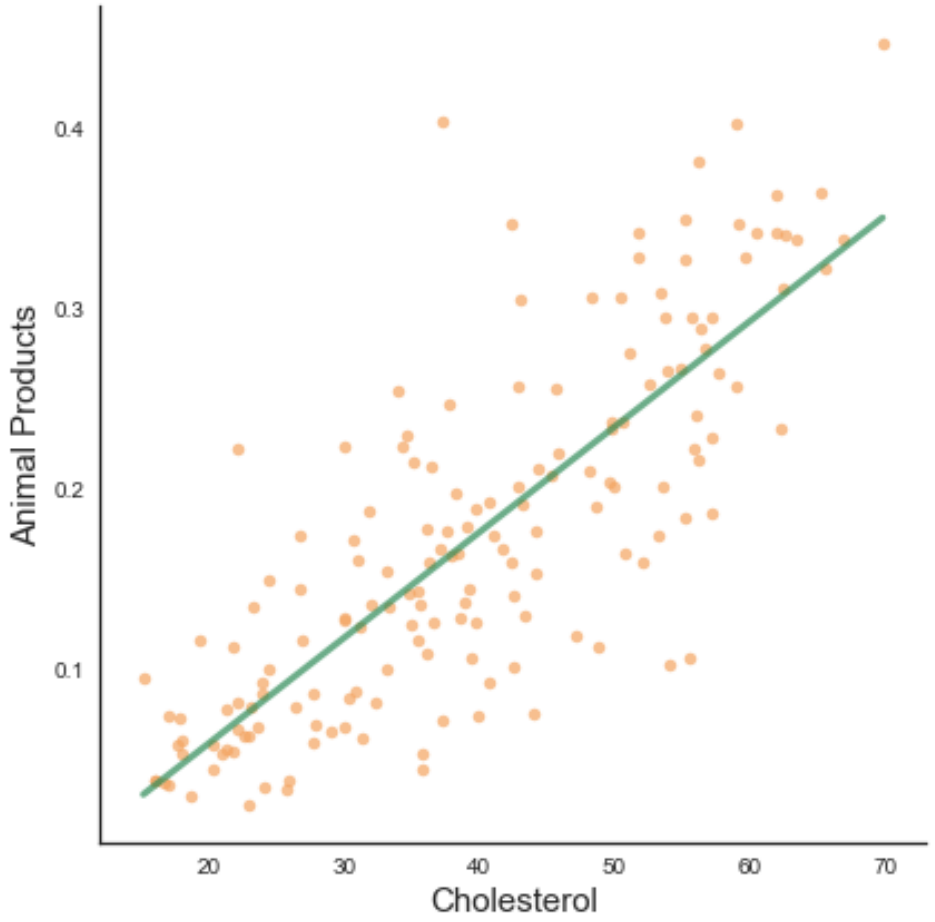


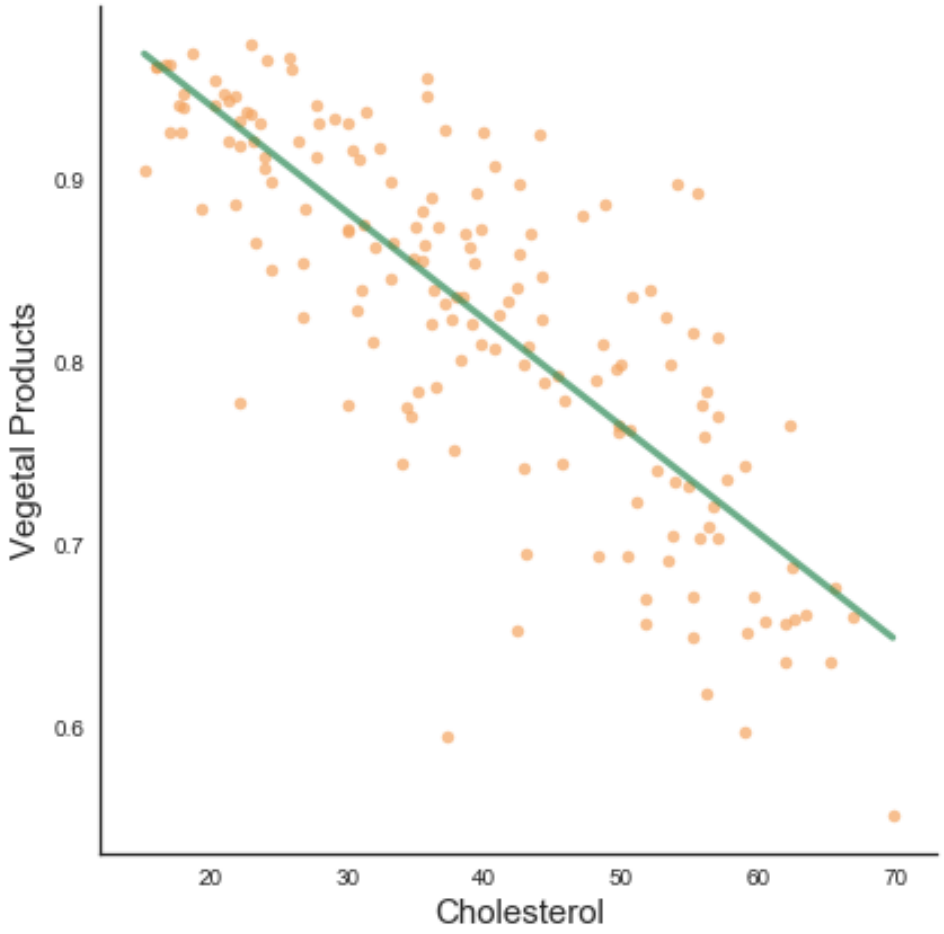
	GNI per capita
Total calories	0.69
Animal products	0.78
Vegetal products	0.22
Animal fats	0.63
Meat	0.56
Offals	0.35
Dairy products	0.61
Eggs	0.62
Fish & Seafood	0.19
Vegetables	0.19
Fruits	-0.04
Cereals	-0.52
Pulses	-0.36
Starchy Roots	-0.26
Sugar & Sweeteners	0.39
Treenuts	0.28
Oilcrops	-0.20
Vegetable oils	0.34



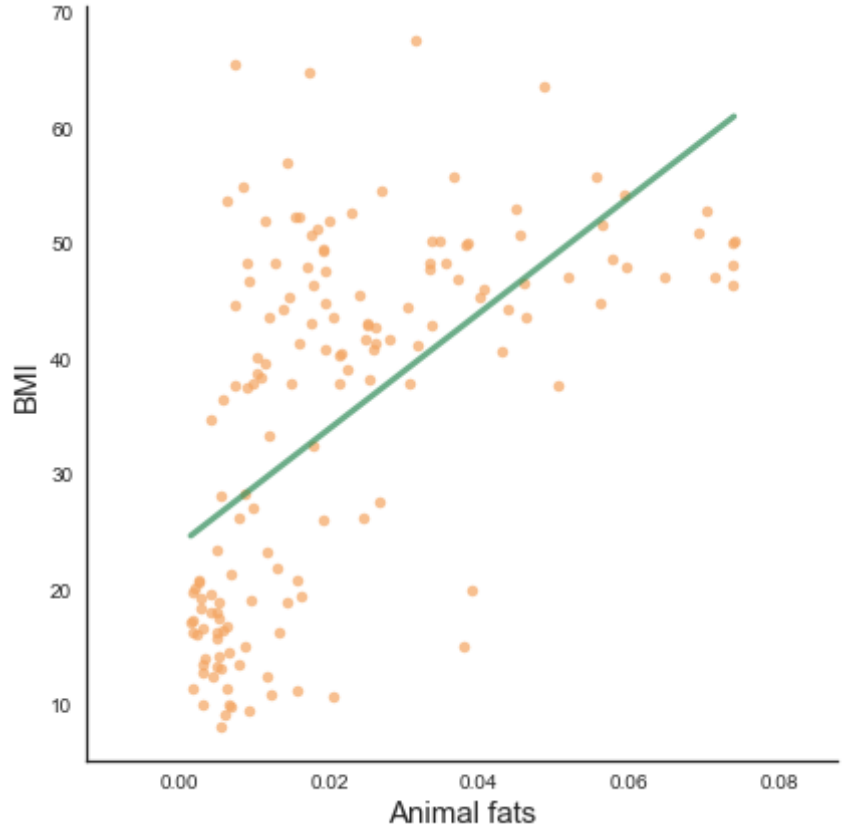


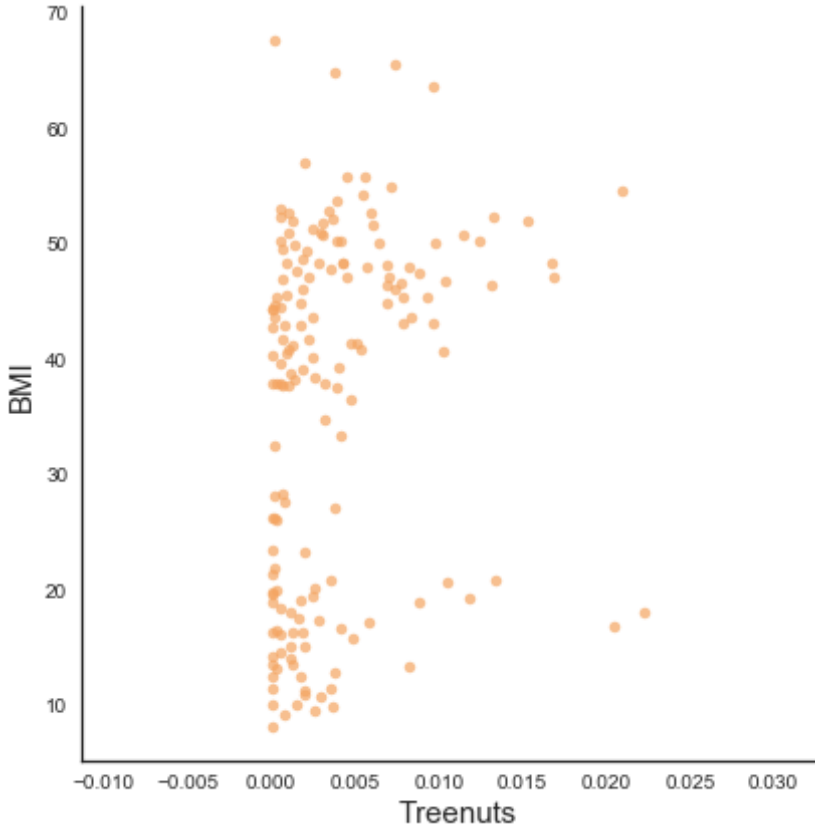
	Animal Products	Vegetal Products
Blood pressure	0.30	-0.30
Blood glucose	0.12	-0.12
Cholesterol	0.80	-0.80
BMI	0.69	-0.69
HALE	0.72	-0.72

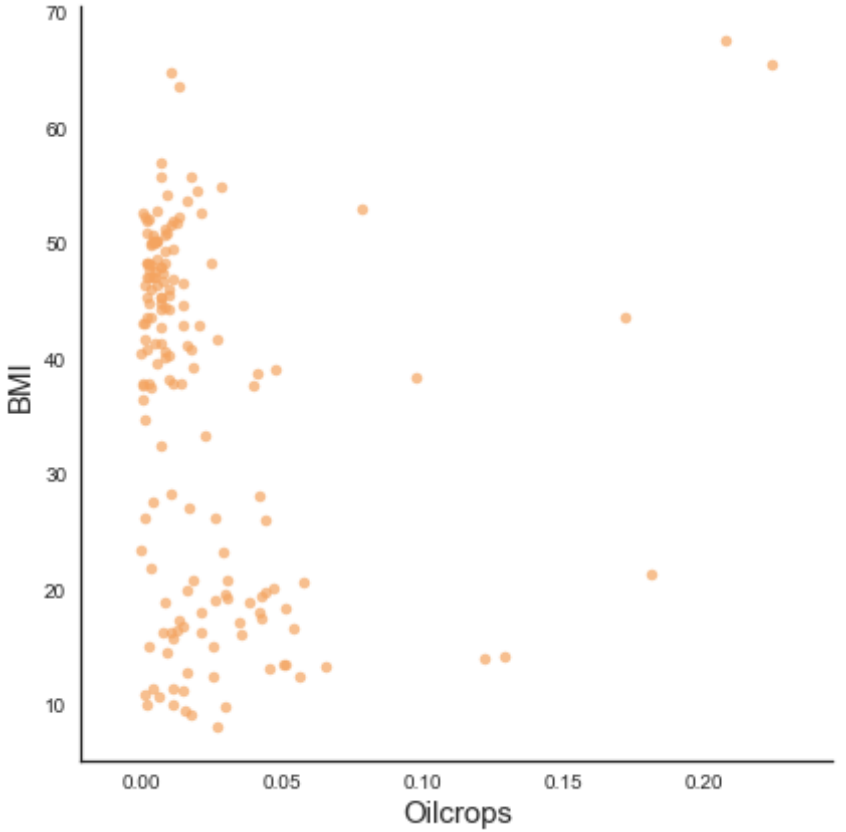




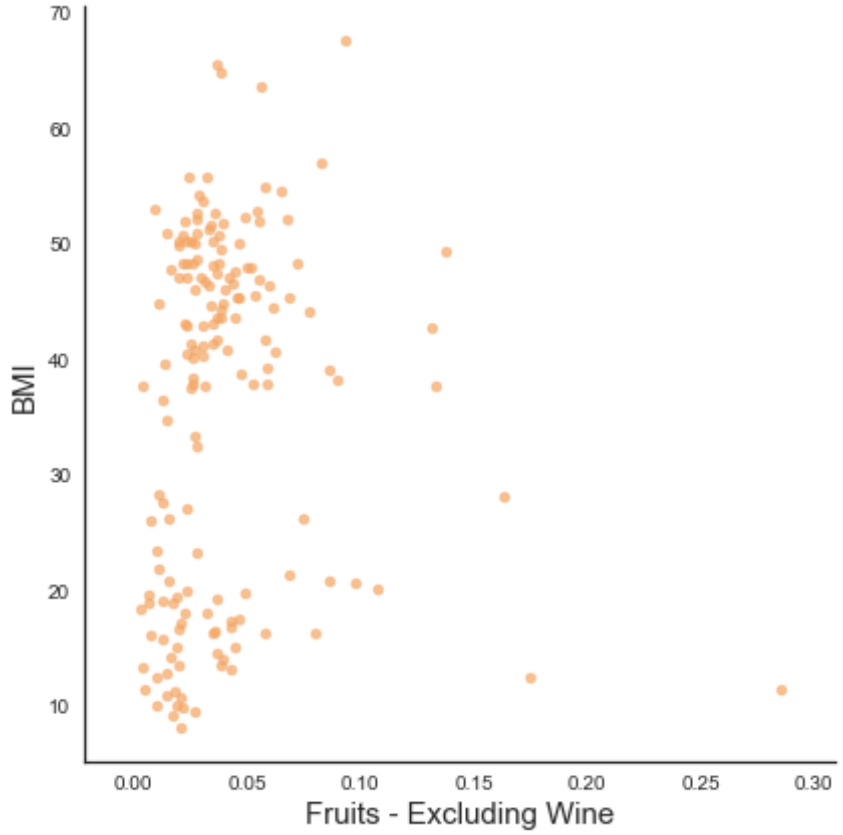
	Animal fats	Meat	Offals	Dairy products	Eggs	Fish & Seafood	Vegetables	Fruits	Cereals	Pulses	Starchy Roots	Sugar & Sweeteners	Treenuts	Oilcrops	Vegetable oils
Blood pressure	0.27	0.22	0.36	0.36	0.18	-0.24	0.11	-0.21	-0.12	-0.20	-0.05	-0.02	0.07	-0.30	0.19
Blood glucose	0.00	0.17	-0.05	0.04	0.07	0.17	0.18	-0.02	-0.04	-0.28	-0.25	0.34	-0.03	0.27	0.02
Cholesterol	0.67	0.65	0.44	0.65	0.79	0.18	0.36	-0.05	-0.51	-0.48	-0.38	0.53	0.26	-0.27	0.29
BMI	0.54	0.59	0.36	0.61	0.62	-0.01	0.37	0.05	-0.53	-0.42	-0.35	0.63	0.22	-0.11	0.33
HALE	0.56	0.56	0.30	0.61	0.76	0.21	0.42	0.00	-0.41	-0.40	-0.48	0.55	0.23	-0.22	0.22

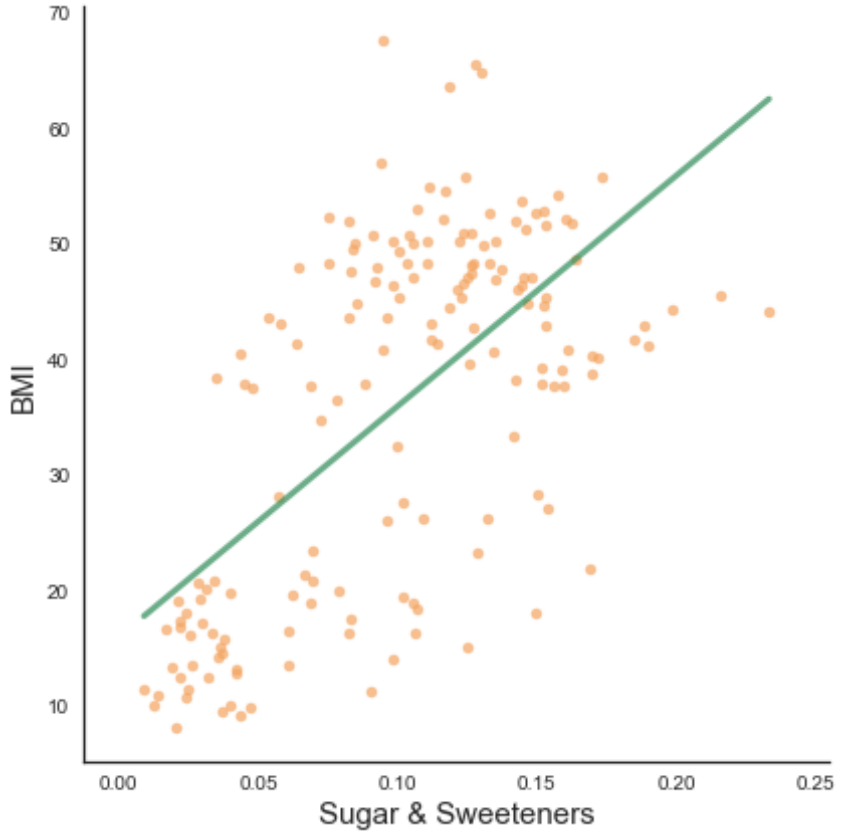






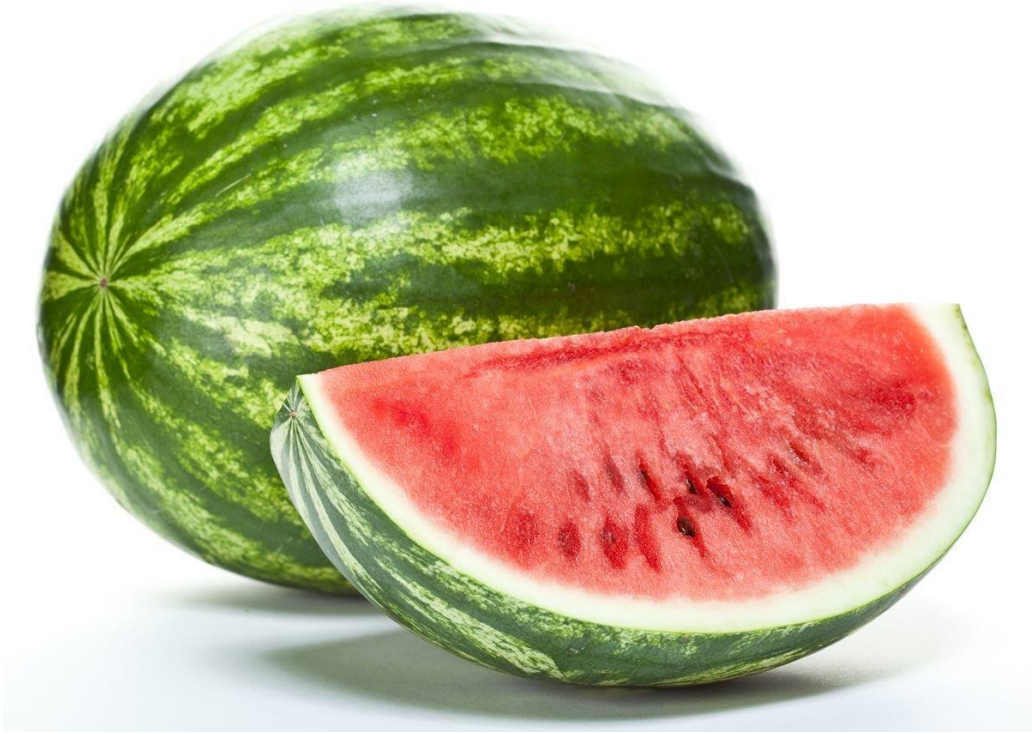
	Animal fats	Meat	Offals	Dairy products	Eggs	Fish & Seafood	Vegetables	Fruits	Cereals	Pulses	Starchy Roots	Sugar & Sweeteners	Treenuts	Oilcrops	Vegetable oils
Blood pressure	0.27	0.22	0.36	0.36	0.18	-0.24	0.11	-0.21	-0.12	-0.20	-0.05	-0.02	0.07	-0.30	0.19
Blood glucose	0.00	0.17	-0.05	0.04	0.07	0.17	0.18	-0.02	-0.04	-0.28	-0.25	0.34	-0.03	0.27	0.02
Cholesterol	0.67	0.65	0.44	0.65	0.79	0.18	0.36	-0.05	-0.51	-0.48	-0.38	0.53	0.26	-0.27	0.29
BMI	0.54	0.59	0.36	0.61	0.62	-0.01	0.37	0.05	-0.53	-0.42	-0.35	0.63	0.22	-0.11	0.33
HALE	0.56	0.56	0.30	0.61	0.76	0.21	0.42	0.00	-0.41	-0.40	-0.48	0.55	0.23	-0.22	0.22





Conclusion

- PCA reduced data managed to preserve clustering of original data.
- Clustering when looking at nutrition and health indicator separately is consistent.
- Be aware of deadly vegetables: Apply caution when using correlation coefficients.
- Nonetheless: Interesting trends may be found
- Let the data tell its story – don't get too caught up in your own narrative.



Thank you for listening!

Sources:

- Food balance sheets:
  fao.org/faostat

- Global Health Observatory:
  who.int/gho/database

- GNI per capita:
  data.worldbank.org

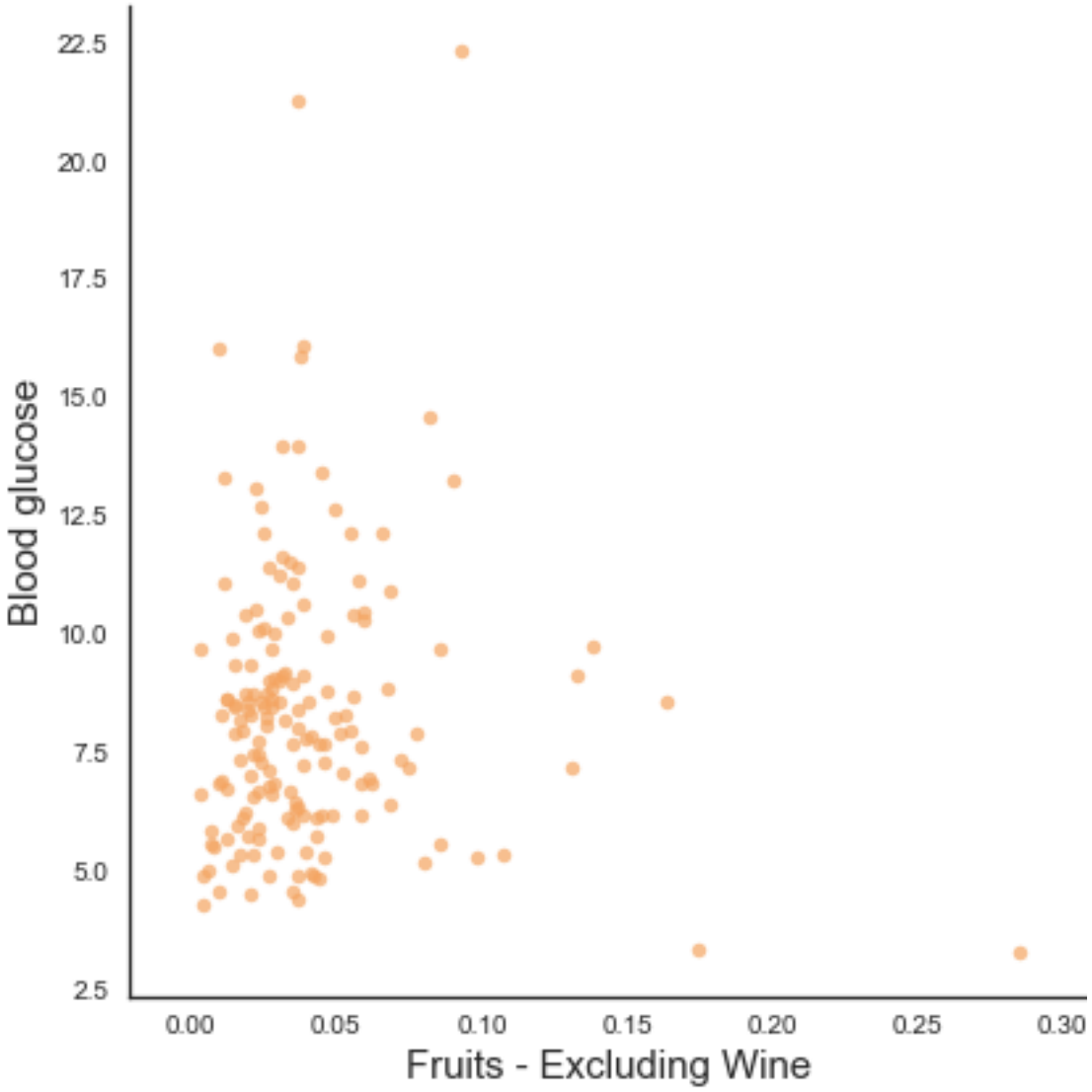
Python libraries used:

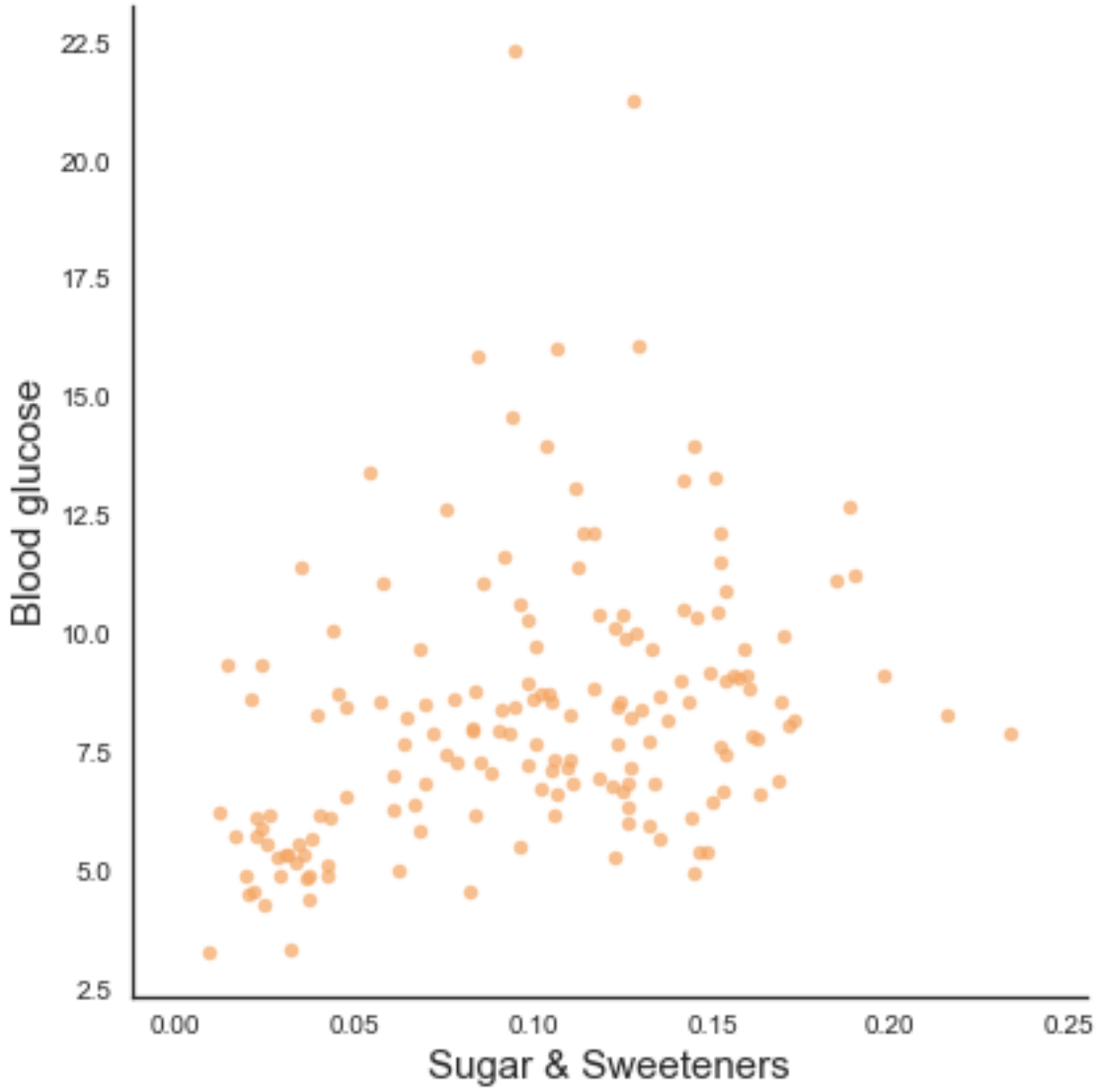
numpy, pylab, pandas, cartopy, seaborn, scipy, sklearn

Interesting visualisation to toy around with:

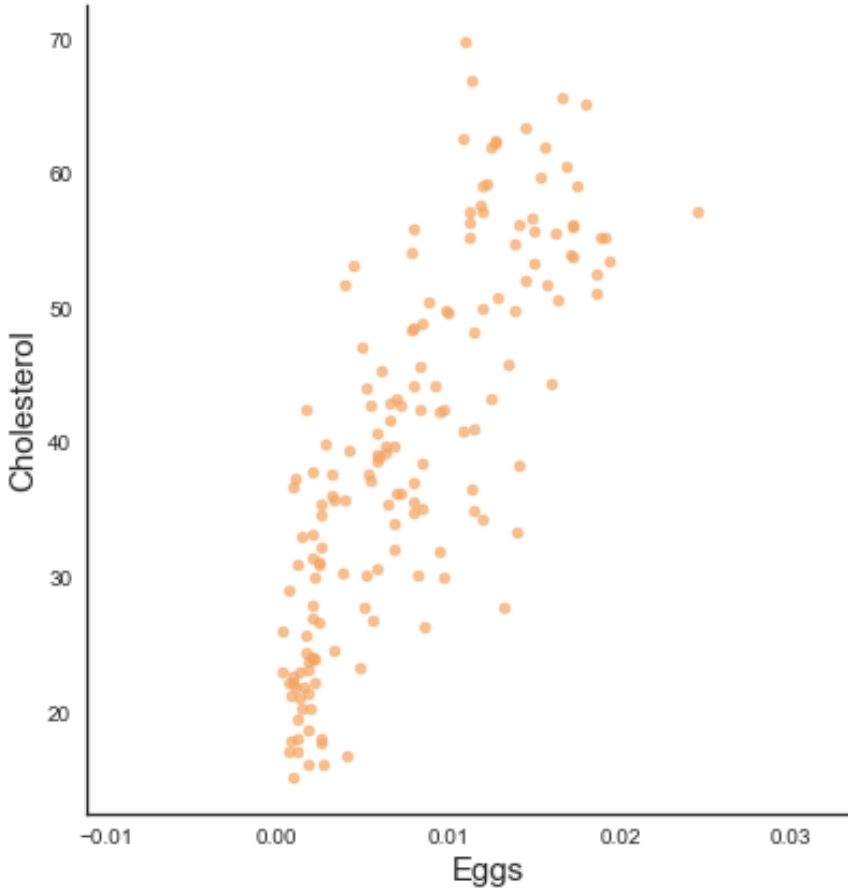
http://www.nationalgeographic.com/what-the-world-eats/

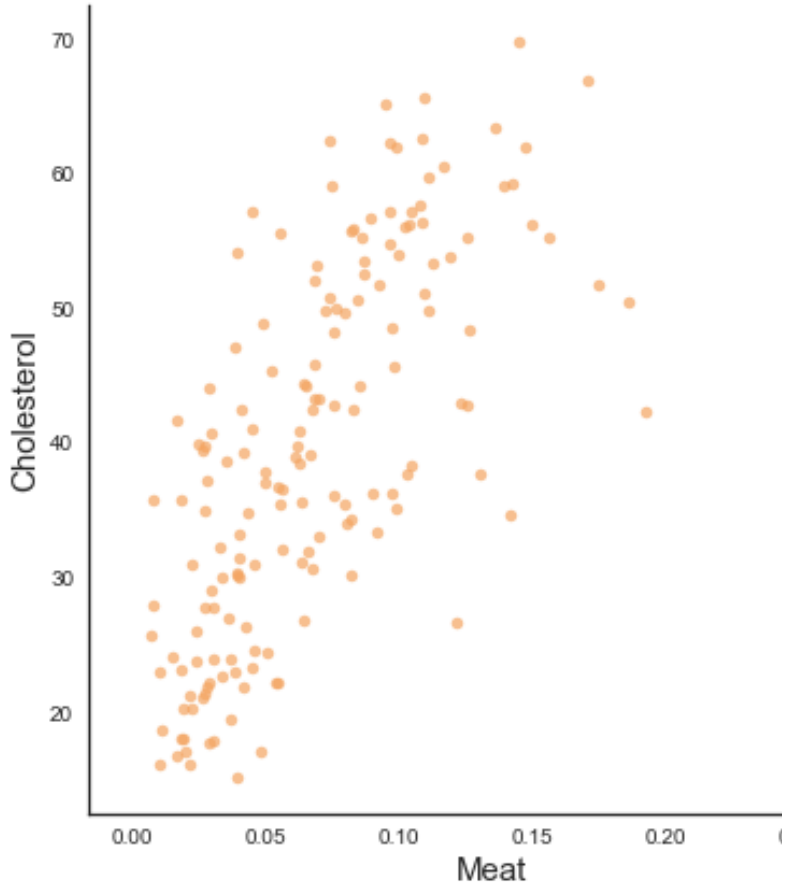
Backup

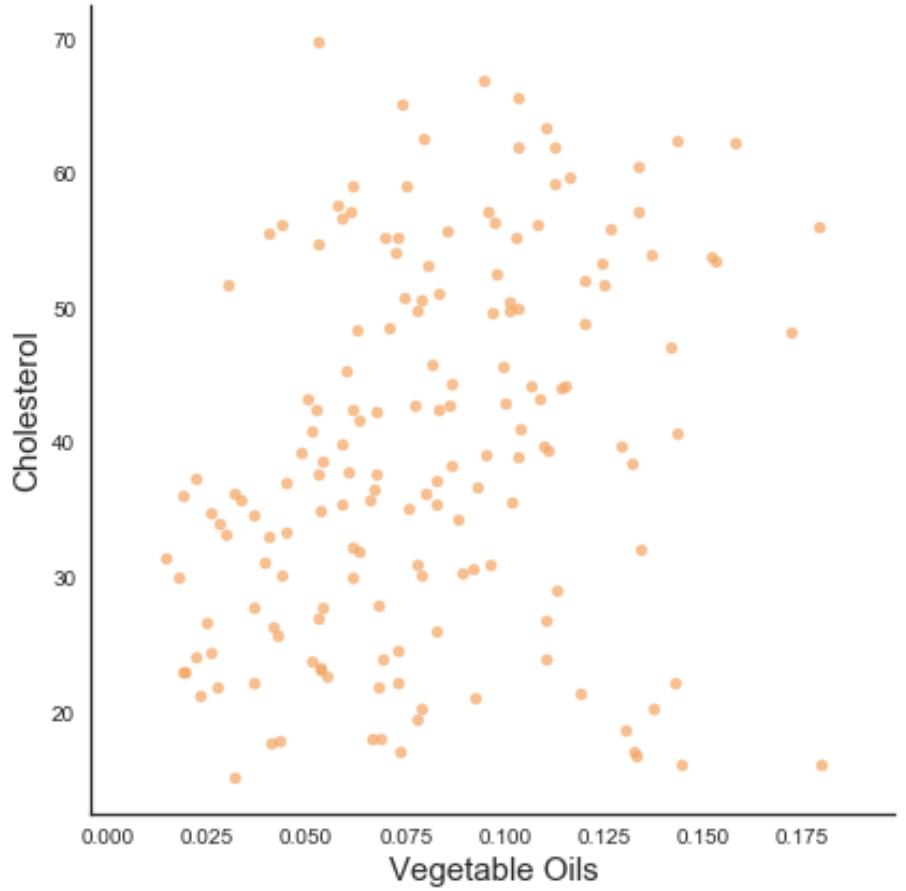

Blood glucose
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Fruits - Excluding Wine


Blood glucose
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.00
0.05
0.10
0.15
0.20
0.25
Sugar & Sweeteners

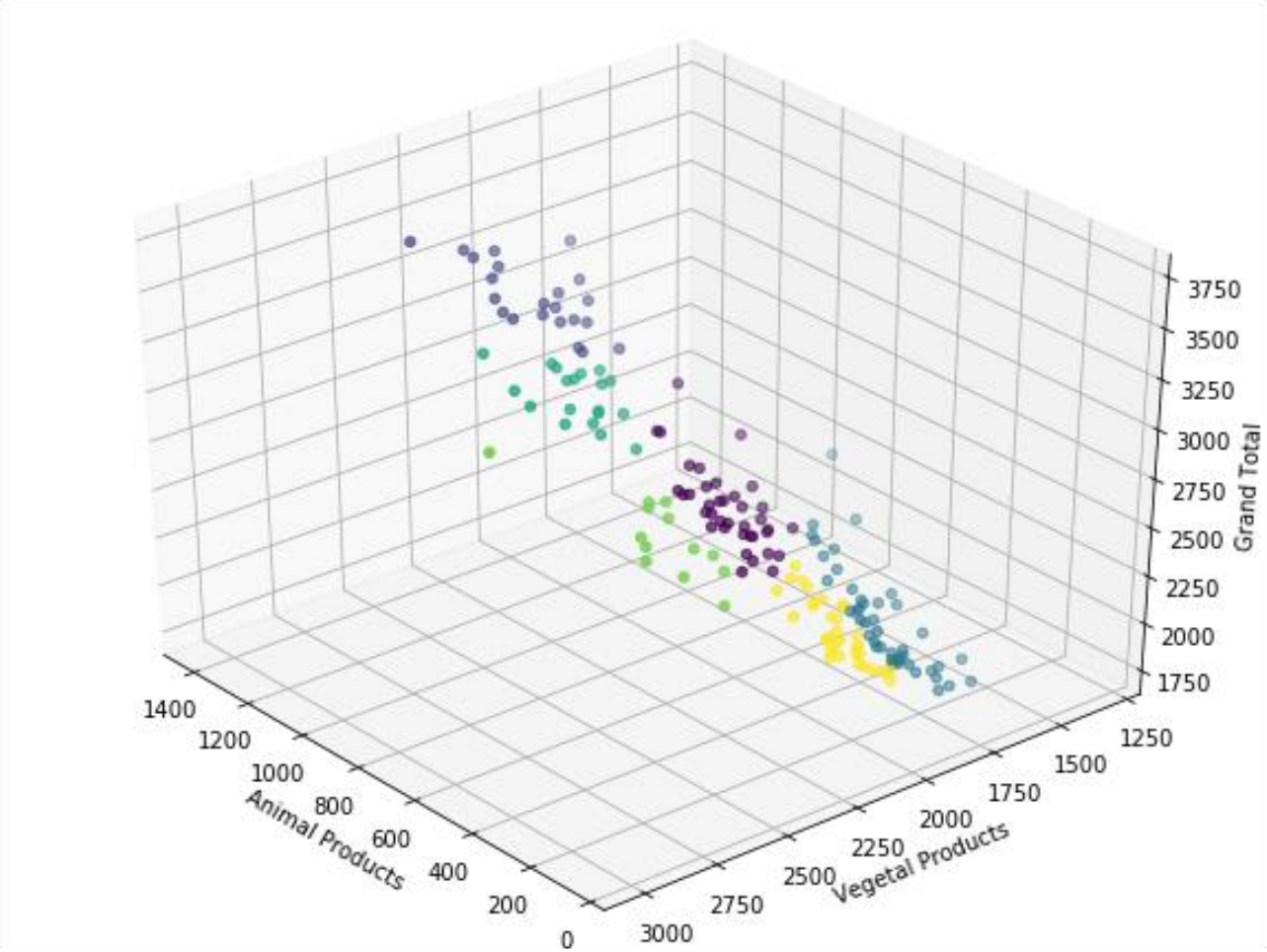
Backup

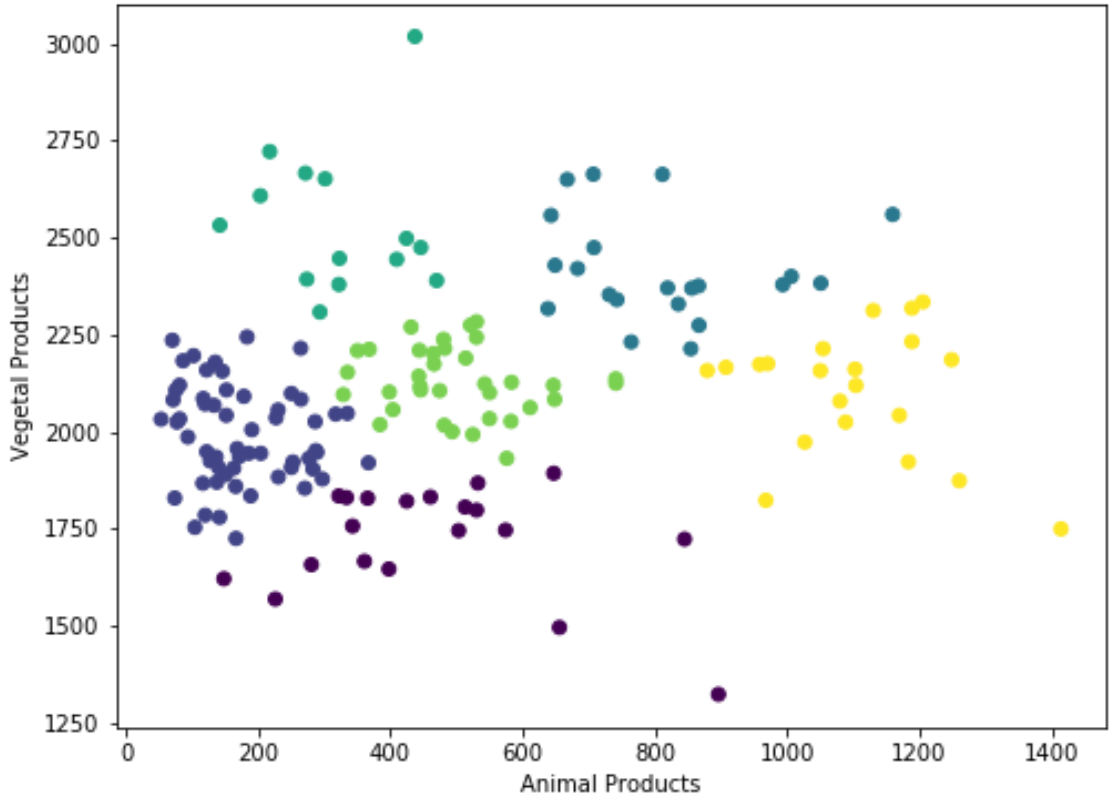

Cholesterol
70
60
50
40
30
20
−0.01
0.00
0.01
0.02
0.03
Eggs


Cholesterol
70
60
50
40
30
20
0.00
0.05
0.10
0.15
0.20
Meat

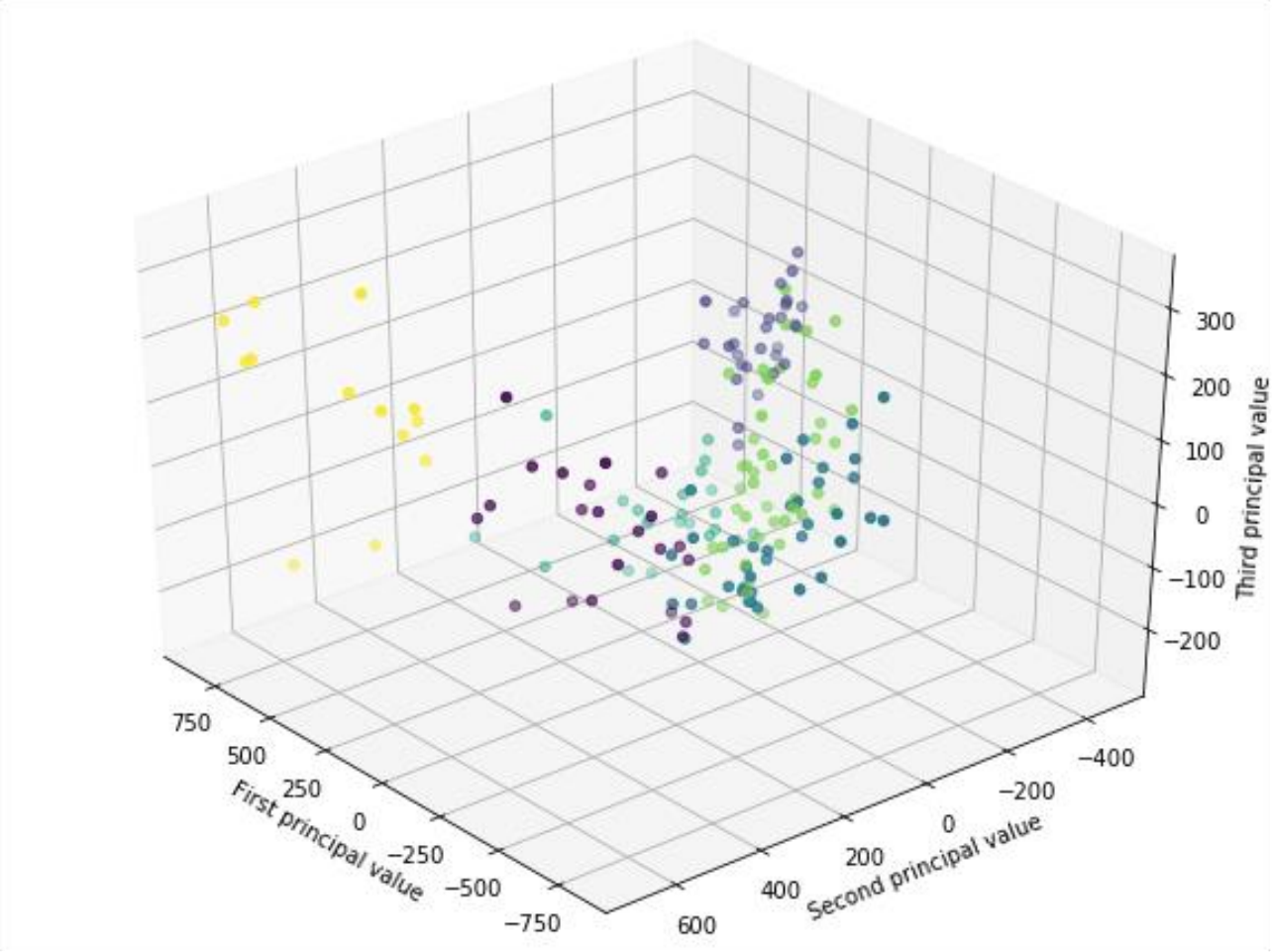

Cholesterol
70
60
50
40
30
20
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Vegetable Oils

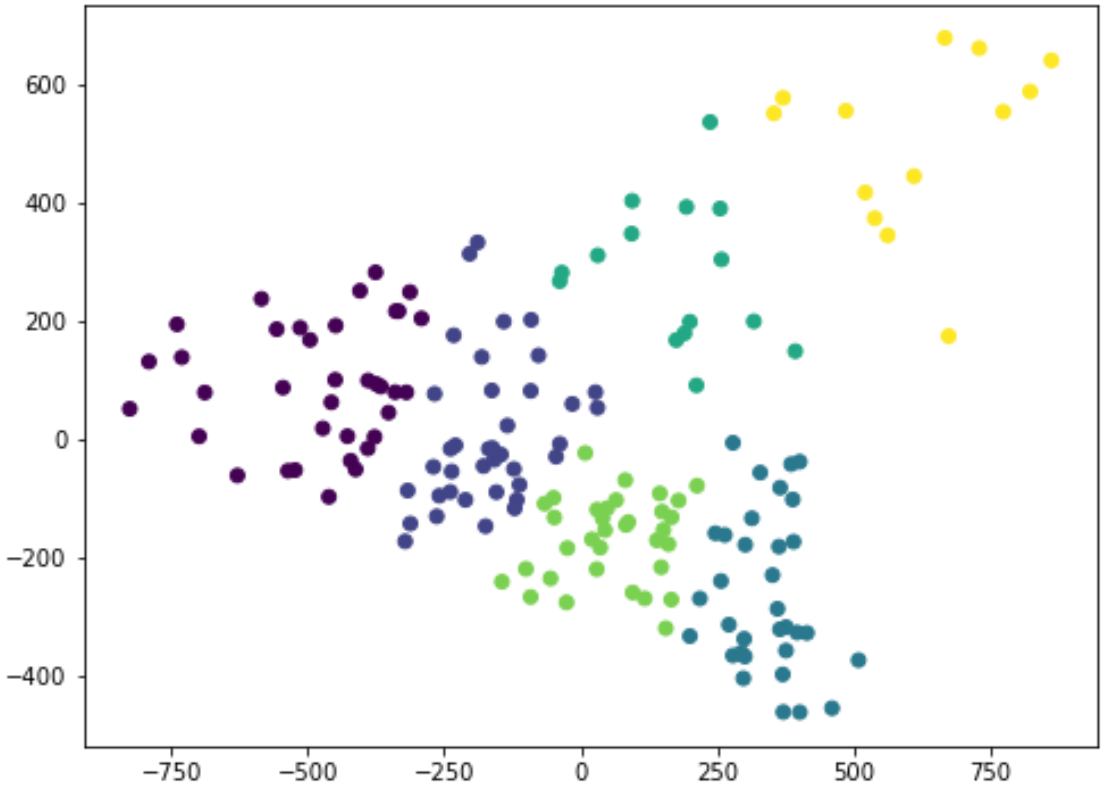
Backup


Animal Products
1400
1200
1000
800
600
400
200
0
Vegetal Products
3000
2750
2500
2250
2000
1750
1500
1250
Grand Total
3750
3500
3250
3000
2750
2500
2250
2000
1750


Vegetal Products
3000
2750
2500
2250
2000
1750
1500
1250
Animal Products
0
200
400
600
800
1000
1200
1400

Backup


First principal value
750
500
250
0
−250
−500
−750
Second principal value
600
400
200
0
−200
−400
Third principal value
300
200
100
0
−100
−200


600
400
200
0
−200
−400
−750
−500
−250
0
250
500
750